# *Trajectories from Mild Cognitive Impairment to Alzheimer's Disease: A machine learning approach in the context of Precision Medicine*

**Ingrid Rye**


**MAPSYK360 Masterprogram i Psykologi,**

**Studieretning: Atferd og Nevrovitenskap**

**UNIVERSITETET I BERGEN**

**DET PSYKOLOGISKE FAKULTET**

**VÅR 2021**

Word count: 14192

*Main supervisor:* Astri J. Lundervold

Department of Biological and Medical Psychology, University of Bergen, Norway

*Co-supervisor:* Alexandra Vik

Department of Radiology, Haukeland University Hospital, Norway

# Abstract

Mild Cognitive Impairment (MCI) is a diagnostic entity including a heterogeneous group of patients. For some, MCI represents a trajectory towards a neurodegenerative disease, while others will remain stable or improve over time. Early identification of a neurodegenerative process is essential to provide treatment before the disease is well established in the brain. This motivated the current study to use longitudinal data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to investigate two groups of patients defined with an amnestic type MCI (aMCI) at a baseline examination: one remaining stable (sMCI) and one converting to Alzheimer's disease (cAD). Variables, selected to represent a proxy to an ordinary clinical examination, included measures of memory and executive function, depressive symptoms, intellectual function, hippocampus volume and ApoE genotype. There were significant differences between the two groups, with the sMCI group showing better performance on tests of memory and executive function, larger volume of hippocampus and fewer ApoE-$\varepsilon$4 positive subjects. We then asked how well a trajectory towards AD could be predicted from the selected variables using a Random Forest (RF) machine learning framework. When evaluated on a test set, the RF model showed a classification accuracy of 68.3%. Computations of feature importance indicated immediate and delayed memory, hippocampus volume and executive function to be most important for this prediction, and partial dependency plots showed cut-off values for increasing risk of conversion. Results are discussed from a clinical, theoretical, and analytic perspective, arguing for their relevance in the context of precision medicine.

***Keywords:*** *Mild Cognitive Impairment; Alzheimer's disease; Neurocognition; Random Forest; Alzheimer Neuroimaging Initiative; Precision medicine.*

# Sammendrag

Mild Kognitv Svikt (MKS) er en diagnostisk kategori som beskriver en heterogen gruppe pasienter. For noen representerer MKS et tidlig tegn på en nevrodegenerativ sykdom, mens andre forbli stabile eller forbedrer seg over tid. Tidlig identifisering av nevrodegenerasjon er svært viktig for å kunne påbegynne behandling før sykdommen allerede har forårsaket store skader i hjernen. Dette motiverte den aktuelle studien, der longitudinelle data fra Alzheimer's Disease Neuroimaging Initiative (ADNI) benyttes for å undersøke to grupper av pasienter som ved baseline viste MKS av den amnestiske typen (aMKS): en gruppe som forble stabile over tid (sMKS) og en gruppe som etterhvert fikk diagnosen Alzheimer's sykdom (cMKS). Det ble valgt ut variabler som gjerne inngår i en klinisk undersøkelse av pasienter med aMKS. Disse omfatter mål på hukommelses- og eksekutiv funksjon, depresive symptomer, intellektuell funksjon, hippocampusvolum og genotype (ApoE). Resultatene viste bedre resultater på tester av hukommelse og eksekutiv funksjon, større hippocampusvolum, og færre individer med ApoE-$\varepsilon$4 i sMKS enn cMKS gruppen. Vi undersøkte deretter hvor godt et utviklingsforløp mot AD kunne predikeres basert på de utvalgte variablene ved å benytte en Random Forest (RF) modell. Evaluering av modellens nøyaktighet i et testset viste en nøyaktighet på 68.3%. Beregninger av de ulike variablenes betydning for klassifikasjonen viste at den var sterkest for mål på hukommelse, hippocampusvolum og eksekutiv funksjon. Partial dependency plots viste terskelverdier som øker sannsynligheten for å klassifiseres i cMKS gruppen. Resultatene diskuteres fra et klinisk, teoretisk og analytisk perspektiv, med vekt på studiens relevans for en fremtidsrettet presisjonsmedisin.

***Nøkkelord:*** *Mild Kognitiv Svikt; Alzheimer's sykdom; Nevrokognisjon; Random Forest; Alzheimer Neuroimaging Initiative; Presisjonsmedisin.*

# Preface

First and foremost I would like to thank my main supervisor Astri J. Lundervold and my co-supervisor Alexandra Vik. Since the project's very beginning a year ago, they have lent me their utmost support and guidance - and they have done so in a socratic manner that has developed my critical thinking skills. Despite some additional challenges due to the COVID pandemic, they have both been there throughout the project's development (although digitally for some periods), open to discuss and give me valuable inputs. I am also very grateful for how they have encouraged me to present my work in forms of abstracts and presentations in several different setting. This has challenged me and facilitated my professional growth.

I would also like to thank the Machine Learning Group at Mohn Medical Imaging and Visualization Centre, with special thanks to Marek Kocinski and Alexander Lundervold. Their collaboration in preparing the scripts for this thesis has been invaluable. They have also given me good advice my deep-diving into the technical world of Python, GitHub and Overleaf. I also want to thank Arvid Lundervold for helping me configure an APA-style template in Overleaf. I am truly grateful I have gotten the chance to take part in such a stimulating and "state-of-the-art" environment.

Finally, data used in preparation of this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. As such, the investigators within the ADNI project contributed to the design, implementation of ADNI and provided data, but did not participate in the analysis or writing of this work.

**Contents**

# 1 Theoretical Background

Impaired cognitive function, and especially problems related to memory, is commonly reported by older adults. When the impairment gets medical attention, the person will in many cases be referred to a memory clinic for a risk evaluation of a neurodegenerative disorder. At the clinic, neuropsychological tests are commonly used to assess function within different cognitive domains (e.g. memory, executive function, and language), and for some, the examination will be extended to include an MRI examination and a blood sample for genetic analysis. The diagnostic label Mild Cognitive Impairment (MCI) will be used if the person shows a cognitive decline that is more severe than expected from her/his age and education level, but still not sufficiently severe to warrant a diagnosis of dementia (Petersen 2004a; Gauthier et al., 2006).

If the phenotypic profile of an MCI patient is defined by a primary memory impairment, it is referred to as an amnestic MCI (aMCI). It is empirically well-established that people with aMCI have a ten-fold increased risk of Alzheimer's disease (AD). However, the cohort of MCI individuals meeting the criteria for this diagnostic entity is immensely heterogeneous both with respect to clinical phenotypes, underlying etiology, and prognostics. Therefore, even though many individuals with aMCI may be on a trajectory towards AD, a substantial proportion of those individuals do not have an underlying neurodegenerative process leading to this disorder, and may never progress to any disorder characterized by dementia. Their symptoms may rather be caused by common treatable conditions like depression, cardiovascular disease, inflammation, and hormone dysregulation (Panza et al., 2018).

Being able to differentiate MCI subjects on a trajectory towards AD from those who remain stable over time or show remission, is a paramount goal in the research field, and for precision medicine more generally. In addition to the obvious clinical importance,

identifying which individuals are on an AD trajectory is of great importance to the success of clinical drug trials. Imagine for instance that a proposed drug in reality is an effective agent in preventing or stagnating progression of AD. The clinical trial of this very drug may nevertheless fail if the study includes a substantial proportion of participants who do not have AD pathology.

To our knowledge, few studies have investigated how well data obtained at the time a patient was first diagnosed with MCI can predict whether this patient will convert to AD. This motivated the present study to investigate characteristics of a group of patients with MCI in an open longitudinal dataset; the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. Two MCI subgroups will be defined based on longitudinal diagnostic status; one including subjects remaining stable with an MCI diagnosis (sMCI) and one group including subjects converting to AD (cAD) throughout their participation in ADNI. The following research questions are raised: Do the two groups differ on selected variables already at an early assessment (baseline), i.e. years before knowing that one of the groups convert to AD? If yes, can this information be used to predict whether an individual will show a sMCI or cAD trajectory, and would it give the clinician knowledge about how to put weight on the different features and their values already at an early visit? To that end, explorative analyses of group differences will be extended by a machine learning approach to investigate the predictions.

Before presenting the methods and results from the empirical study, a theoretical background for the selection of themes, variables, and statistical approaches will be presented in the following section.

## 1.1 From normal to pathological aging

Questions related to how aging affects brain function have interested scientists for decades. Many elderly will experience minor glitches in memory. While some will let

them slide thinking that they are "just part of getting old", for others these same glitches may lead to concerns that heavily impair daily life functioning. Today there is a broad consensus among experts in the field that some cognitive abilities, such as verbal knowledge and semantic memory increases across the lifespan, whereas other abilities including processing speed, working memory and episodic memory consistently show decline with age (Park et al., 2002; Oh et al., 2012).

We see, however, that the cognitive changes associated with aging are characterized by diversity in phenotype, with respect to both pace and severity. This diversity is a result of the several biological and lifestyle-dependent factors influencing an individual throughout the lifespan (Walhovd et al., 2014, Nyberg, 2019). Individuals who preserve their cognitive function into old age are found at the one end of a continuum of cognitive aging, including elderly referred to as "superagers" (Rogalski et al., 2013). At the other end, we have individuals who may experience cognitive decline at a much younger age due to neurodegenerative disease (Petersen et al., 2006). Along this wide dimension of cognitive function, it becomes difficult to define the fine distinction between normal and pathological aging, and to predict a trajectory towards a neurodegenerative disorder from clinical signs at an early stage of the disease. In the present thesis, the focus will be on the trajectory from such early signs of impairment towards Alzheimer's disease (AD), one of the many disorders associated with dementia.

### *1.1.1 Alzheimer's disease (AD)*

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder estimated to cause around 60-90% of all cases of dementia (Huang et al., 2020; American Psychiatric Association [APA], 2013). The disease is typically divided into early- and late-onset AD, distinguished by age at onset with 65 years old being the cut-off (Reitz et al., 2020). To obtain a diagnosis of AD, the cognitive impairment should have an insidious onset, and be se-

vere enough to interfere with functions of daily living (APA, 2013). Impairment of episodic memory function is the most common initial symptom of the disorder. This typically manifests as forgetting recent events and conversations, as well as problems with learning new information. Then follows progressive decline within other cognitive domains, often accompanied by alternations in emotional control, motivation and social behavior. As the disease advances, the patient will gradually lose his/her ability to complete basic daily life activities such as eating, dressing and personal care. As of today, there are no treatments available to revert or cure the disease, and the average duration of dementia due to AD is estimated to 7-10 years with death as an inevitable endpoint (Holtzman et al., 2011).

AD is posing a major challenge in todays society and it is recognized as a major epidemic (Hampel et al., 2011; Sperling et al., 2011). With increased longevity, the elderly proportion of the population grows, and with age being the primary risk factor for AD, the global community is facing great challenges related to the disease in the coming years (Winblad et al., 2016). Alongside the devastating personal consequences a diagnosis of AD has on those affected and their caregivers, the economical costs are massive. In a report published in 2019, the current economical costs associated with AD in Norway were estimated to constitute 62 billion NOK (Menon Economics, 2020, p. 19). The report further outlines a prospective analysis concluding that without new and effective treatments to cure or stagnate the progression of AD, the costs related to the disease will almost triple (estimated to 180 billion NOK) within the year 2040. Comparable estimates are foreshadowed globally (Prince et al., 2015).

Effective treatment for AD is therefore strongly called for. Today, the field is challenged by problems related to early detection. It is well established that the degenerative process of AD starts decades before the clinical signs. When these signs are severe enough to get medical attention, extensive neural degeneration is already well established in the brain (Braak & Braak, 1991). This fact has lead to intensive research in the field focusing

on predementia stages of neurodegenerative diseases. As already stated, memory problems are reliable signs of AD. These signs are, however, difficult to distinguish from memory problems frequently reported by older adults. An extensive examination should therefore be conducted to identify prodromal signs of AD, with specific memory tests to identify 'true' AD-related memory impairment, in addition to tests assessing other cognitive domains (Dubois et al., 2009).

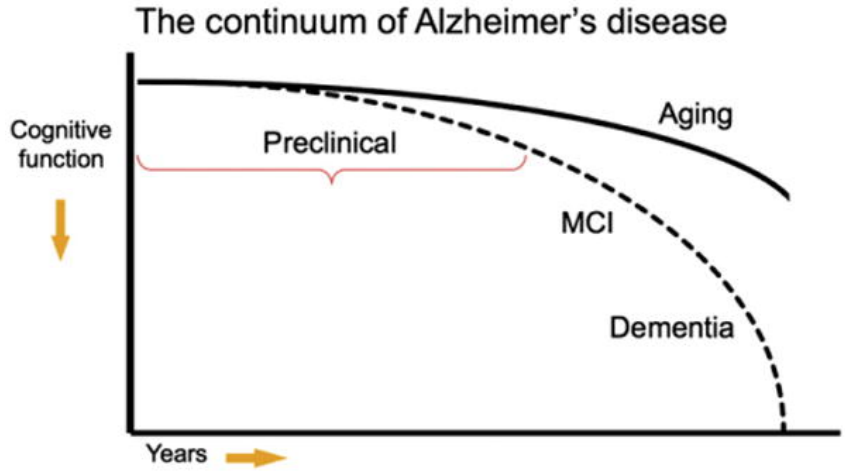
Taken together, longitudinal studies identifying cognitive changes associated with different steps from normal cognitive aging to AD and other neurodegenerative disorders are indeed called for. They are important to enable early identification and treatment, but also to identify characteristics of changes experienced by a patient along a trajectory with a given outcome. As a response to this call, Petersen and colleagues (1999) introduced the diagnostic construct of MCI to describe the transitional stage between normal cognitive function and dementia. Individuals falling within this diagnostic category have a cognitive decline greater than what is expected for normal aging, but the impairment is still not severe enough to warrant a diagnosis of dementia (i.e. activities of daily living are mainly preserved) (see Figure 1).

### *1.1.2 Mild Cognitive Impairment (MCI)*

The first official criteria for MCI was formulated by a group of researchers at the Mayo Clinic and was originally intended to capture individuals with prodromal AD (Petersen et al., 1999). To obtain a diagnosis of MCI according to these original criteria, the patient had to have memory complaints which could also be corroborated by objective deficits on tests of episodic memory. Importantly however, impairments should not be severe enough to warrant a diagnosis of dementia. With an increasing amount of studies employing these MCI-criteria being published, it soon became clear that a substantial proportion of patients defined as MCI never progressed to AD. It was therefore decided

**Figure 1**

*Model of the clinical continuum of Alzheimer's disease. Illustration of cognitive decline as a function of normal (solid line) and pathological (dotted line) aging. Figure adapted from Sperling et al. (2011).*




that the diagnostic construct of MCI needed to be broadened to encompass this heterogeneity. On an international consensus conference held in 2003, the original criteria from the Mayo Clinic were thus expanded to encompass cognitive impairments affecting cognitive domains other than memory (Winblad et al., 2004).

In 2011, a working group from the (American) National Institute on Aging and Alzheimer's Association met to discuss the criteria for the symptomatic predementia phase of AD. At that meeting, they proposed a more specific definition of 'MCI due to AD' (Jack et al., 2011). According to their diagnostic guidelines, four core clinical criteria should be fulfilled for a patient to receive a diagnosis of MCI to be obtained: i) a subjective concern regarding change of cognition reported either by the patient, an informant who knows the person well, or a clinician; ii) objective impairment in one or more cognitive domains; iii) generally preserved independent function of daily living; and lastly iv) the patients should not meet the criteria for a diagnosis of dementia.

Despite extensive research and several revisions of diagnostic criteria over the last

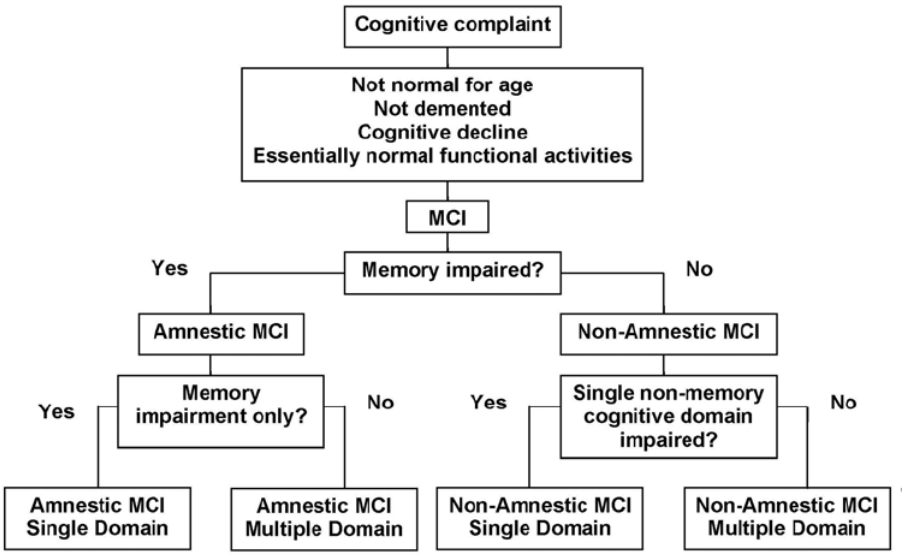
two decades, the MCI construct remains a topic for discussion. The persistent lack of consensus is illustrated by changes incorporated into the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) published in 2013 (APA, 2013). In this newest version, the diagnostic category previously referred to as dementia was replaced by a chapter entitled 'Neurocognitive Disorders' (NCD). The chapter is further differentiated into 'minor' and 'major' NCD, two grades of severity distinguished by whether or not the cognitive decline is severe enough to compromise daily function. Whereas major NCD, when etiology is known, is coded as subtypes (e.g. due to AD, Lewy-body or frontotemporal), etiology for minor NCD is not coded. The inclusion of minor NCD is meant to capture those in a predementia state. It is clearly overlapping with the construct of MCI, with an intent to reflect the emerging literature on this topic (Blazer, 2013).

As research on the MCI cohort has developed, awareness of the heterogeneity characterizing the diagnostic entity of MCI has improved. It has become evident that for some patients, treatment of other diseases may revert the MCI symptoms, others will remain stable over time, and the rest will experience a trajectory towards a neurodegenerative disorder. An initial attempt to tackle this heterogeneity was made by Petersen (2004a) almost 20 years ago, when the first comprehensive clinical stratification of MCI subtypes was presented. Following this nosology, an important distinction is made between amnestic (aMCI) and non-amnestic (na-MCI) subjects with MCI, in which the former group primarily presents with memory impairments whereas the latter group is characterized by an impairment within cognitive domains other than memory. These two groups are further divided into single- or multi-domain types, based on whether the patient's impairment is isolated to one cognitive domain or whether several domains are affected (See Figure 2). According to this stratification, a person with a clinical picture characterized by memory deficits accompanied by preserved cognition in other domains are classified as "single-domain aMCI", whereas a person with intact memory,

but impaired executive function and language problems is classified as "multi-domain na-MCI".

**Figure 2**

*Algorithm for stratification of MCI subtypes. Figure adapted from Petersen (2004b).*




Among the aforementioned subtypes, patients in the non-aMCI subgroup are more likely to progress to a non-AD neurodegenerative disease, like dementia due to Lewy bodies, frontotemporal dementia, or vascular dementia (Peterson, 2004b; Molano et al., 2009), whereas the aMCI type is associated with the highest risk of progression to AD. The estimated rate of progression among these patients is estimated to 10-15% per year, which is considerably higher than in the general population of older adults, progressing at a rate of 1-2% per year (Liu et al., 2013). aMCI is, however, also frequently found among patients with neuropsychiatric disorders like depression as well as other somatic diseases. Although these patients tend to be more stable than the ones progressing to AD, it is often difficult to differentiate between these two groups.

In the present thesis, the focus will be on patients defined as aMCI. It is therefore

important to emphasize that more recent research has uncovered substantial heterogeneity both in cognitive profiles and patterns of atrophy even within this amnestic subgroup. This was for instance illustrated by a study conducting a cluster analysis including 825 individuals defined as aMCI from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Edmonds et al., 2014). Based on subjects' performance on neuropsychological tests covering three domains of cognition (memory, attention/executive function, and language), four empirically derived subtypes were identified. The four subtypes were named: Dysnomic; Dysexecutive, Amnestic, and Cluster-Derived Normal. The last group was especially surprising in that individuals in this group performed within normal limits on the cognitive tests, despite being defined as aMCI patients by the conventional diagnostic criteria used in ADNI. This "misclassification" was found in more than one-third (34%) of the aMCI sample.

The authors drew two main conclusions from this finding. Firstly, they argue that the empirically derived identification of MCI subtypes within this group of aMCI demonstrates a heterogeneity in the cognitive profiles of aMCI patients and that this diversity is not captured by conventional diagnostic criteria. Secondly, they claim that their study indicates a weakness with the conventional diagnostic criteria used for aMCI, with vulnerability to false positives. A follow-up study (Edmonds et al., 2016) on the same cohort further illustrated that the Cluster-Derived Normal subgroup had normal cortical thickness at baseline despite being defined as aMCI. They further found that subjects in this subgroup continued to show normal cognition and minimal cortical atrophy over the next 3 years.

Another similar study identified four atrophy subtypes in an AD sample and retrospectively illustrated that these subtypes could be detected already in the prodromal phase (ten Kate et al., 2018). They further robustly replicated their findings across three independent data sets, giving additional confidence in that the findings indeed reflect true pathophysiological subtypes of AD and its prodromal stage. Findings such as these underscore

the importance of more detailed investigations of predictors for conversion from aMCI to AD type dementia, including both information about cognitive function and biomarkers characterizing patients with aMCI.

### *1.1.3 Biomarkers of MCI and AD*

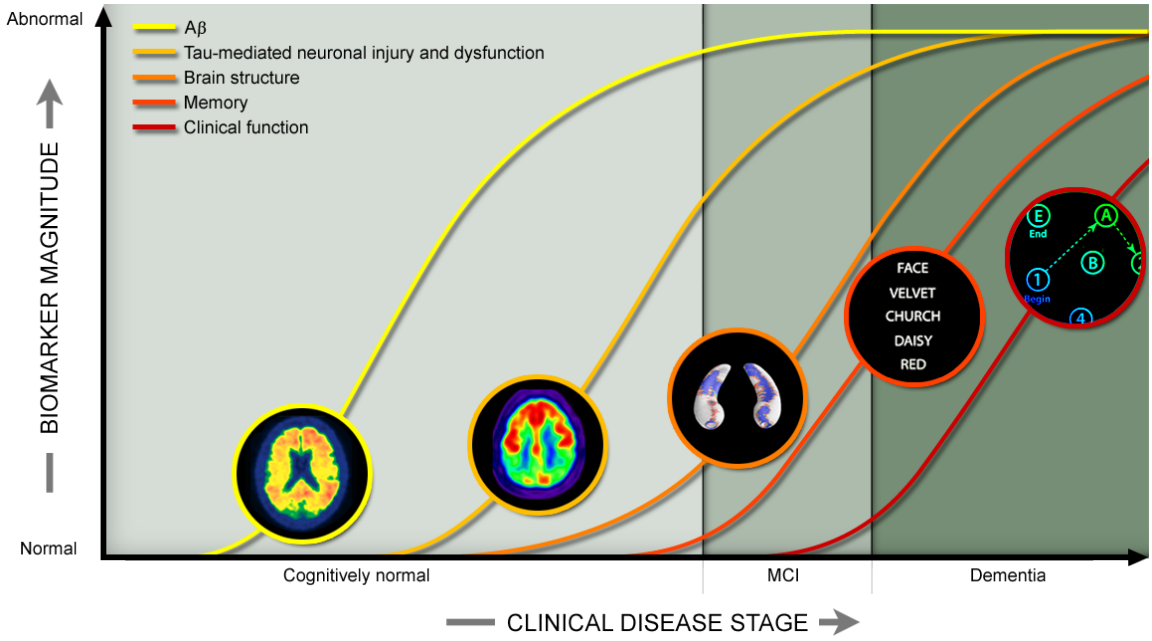
The pathological confirmation of AD requires presence of amyloid beta ($A\beta$) deposition in plaques along with evidence of tau tangles (Albert et al, 2011), and it is this characteristic proteinaceous pathology that differentiates AD from other forms of dementia including, but not limited to, dementia due to Lewy Bodies, frontotemporal dementia, and vascular dementia.

Historically, a definite diagnosis of AD required post-mortem inspection of brain tissue to confirm evidence of AD pathology. Today measures of such biomarkers can be used to increase certainty about etiology and underlying pathology, and as such guide differential diagnosis in living patients. If a patient fulfills the clinical criteria for dementia and the presence of AD biomarkers is confirmed, a probable or possible diagnosis of dementia is given, depending on the degree of certainty (APA, 2013; Gutches, 2019). There is a consensus among most experts in the field of AD that the pathology associated with AD exists on a continuum resulted from a process evolving several decades prior to the manifestation of clinical symptoms (Petersen et al., 2009). As illustrated in Figure 3, AD pathology should thus also be present in individuals with MCI who are on a trajectory towards AD.

Even though the exact mechanisms and order by which the pathology manifests, as well as how it relates to cognitive impairments, are still largely unknown (Jack & Holtzman, 2013), both senile plaques of $A\beta$ and neurofibrillary tau tangles are known to interact and alter synaptic plasticity, leading to synaptic loss, dysfunctional neural network, and eventually neuronal loss (Ricciarelli & Fedele, 2017). A thorough review of the cellular

**Figure 3**

*The graph demonstrates a model for the temporal changes of biomarkers along the cognitive continuum from healthy to Alzheimer's disease. Illustration adapted from ADNI, n.d., (http://adni.loni.usc.edu/study-design/).*

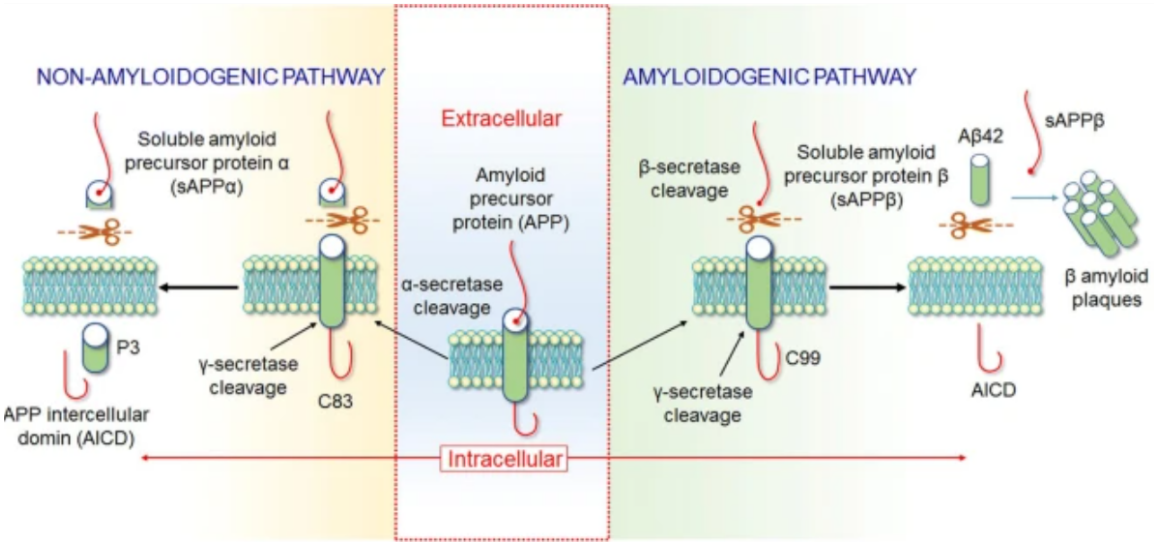



mechanisms involved in these processes is beyond the scope of this thesis (see Calabrò et al., n.d.), but a general overview and how it relates to the biomarkers used in this study will be provided in the following.

***Amyloid beta plaque and neurofibrillary tau tangles.*** To understand the pathophysiology associated with abnormalities in amyloid beta (A$\beta$), it is necessary to understand normal function. The neuronal cell membrane consists of numerous proteins, including a protein called amyloid precursor protein (APP). APP plays an important role in neuronal growth and repair after injury, and as with all proteins in the body, it eventually needs to be recycled and resynthesized. This breakdown happens mainly through two pathways; the non-amyloidogenic pathway and the amyloidogenic pathway (Rhaman et al., 2020) (see Figure 4).

**Figure 4**

*Illustration of the amyloidogenic and non-amyloidogenic pathway for cleavage of amyloid precursor protein (APP). Illustration adapted from Rahman et al. (2020).*




Through the former pathway, APP is broken down by the enzymes alfa-secretase ($\alpha$-secretase) and gamma-secretase ($\gamma$-secretase), which results in the formation of smaller, soluble peptides which are further metabolised and cleared from the extracellular space. In the amyloidogenic pathway however, another enzyme called beta-secretase ($\beta$-secretase) works with $\gamma$-secretase to cleave APP. Through this pathway, the cleavage happens at another location of APP, which results in slightly different peptides called A$\beta$ monomers. Due to the biochemical properties of A$\beta$ monomers, they are insoluble and therefore harder to clear from the extracellular space. Instead, many of these monomers will aggregate in the synaptic junction, initially forming amyloid oligomers, which further aggregates to form senile plaques of A$\beta$.

The other pathological hallmark characterizing AD is neurofibrillary tangles of the tau protein. The primary physiological function of tau proteins is to stabilize the axonal microtubule, an important part of the cell's cytoskeleton (Calabrò et al., n.d.). The micro-

tubule extends from soma to the axon terminal, giving the neuron it's structure and facilitates transport of molecules. In AD, an abnormally large proportion of tau proteins become phosphorylated. In this phosphorylated state, the tau proteins detach from the microtubule and instead cluster together forming neurofibrillary tangles, resulting in the breakdown of microtubules (Iqbal et al., 2005). The intracellular tau tangles disrupt neuronal signaling and eventually lead to cell death causing the neural degeneration characteristic of AD. At the microscopic level, the degeneration is characterized by neuronal loss and at the macroscopic level, it is observed as atrophy (i.e. loss of brain tissue) (Jack & Holtzman, 2013). Consistent with impairments in episodic memory being the initial clinical presentation of typical AD, the spatio-temporal pattern of progression for neurofibrillary tangles in AD subjects start in the transentorhinal cortex, spreads to the hippocampus, and then progresses to cover the cerebral cortex in later stages (Braak & Braak, 1991; Serrano-Pozo et al., 2011) (Figure 5).

***The ApoE gene.*** The ApoE gene is identified as the main genetic risk factor for developing late-onset AD (Liu et al, 2013; Berkowitz et al., 2018), estimated to account for 27.3% of the risk of developing the disease (Van Cauwenberghe et al., 2015). The gene is closely related to the aggregation of A$\beta$ and the tau-related pathology associated with AD (Butt et al, 2021), and information about this gene is therefore included in the present study as a proxy for the biomarkers mentioned above. The ApoE gene codes for Apolipoprotein E, a protein playing a pivotal role in the transport and metabolism of plasma proteins, including APP. There are three isoforms of the ApoE gene; $\varepsilon$2, $\varepsilon$3, and $\varepsilon$4, and it is well-established through both animal (Castellano et al., 2011) and human studies (Roda et al., 2019) that the different isoforms differentially affect both production and clearance of A$\beta$ (Liu et al., 2013). In general, carriers of ApoE-$\varepsilon$4 tend to show lower performance on cognitive tests than non-carriers (Wisdom et al., 2011), and several studies have documented a high prevalence of $\varepsilon$4 alleles among individuals with MCI (Tervo et al., 2004;

Kryscio et al., 2006). Studies assessing ApoE status in relation to fluctuations from MCI to cognitively normal have found the presence of at least one $\varepsilon 4$ allele to be negatively associated with reversion to normal cognition (Koepsell & Monsell, 2012). It is also widely shown that individuals with MCI who are carriers of the $\varepsilon 4$ allele are at increased risk for progressing to AD-type dementia (Xu et al., 2012; Samaranch et al., 2011).

***Brain atrophy and hippocampal volume.*** Although it is well known that the brain changes as we get older, the course of the aging brain is still very much an enigma. The last 30 years of neuroimaging research using Magnetic Resonance Imaging (MRI) has, however, significantly improved our understanding of how the brain changes as we age. Morphometrical studies of the aging brain, e.g. frontal lobe atrophy, hippocampal shrinkage, cortical thinning, ventricular enlargement, can be described as the first 'imaging era' in this field. After the introduction of diffusion tensor MR imaging came the loss of white matter integrity approach to aging, and BOLD fMRI with and without-a-task has enabled assessment of functional aspects of the aging brain. And recently, combining these techniques into *brain connectivity mapping* has moved the field towards a system approach to brain aging (Raz & Kennedy, 2009).

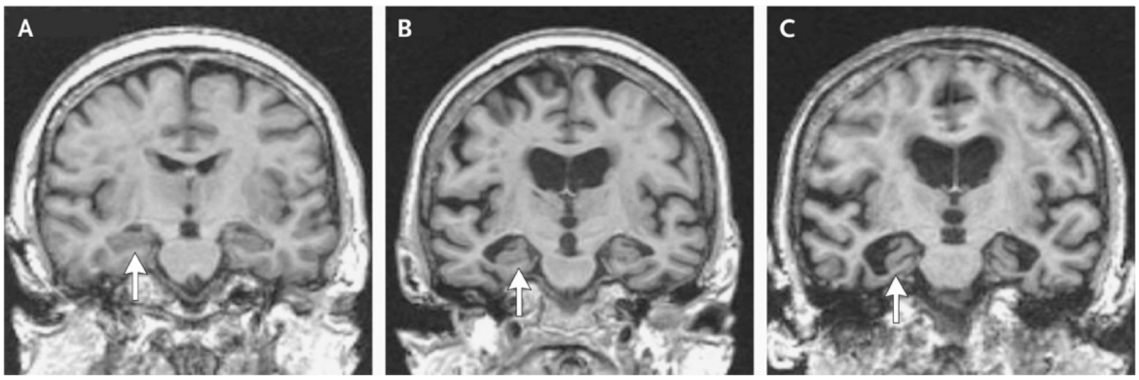
In the present thesis, information about the brain is restricted to a measure of the hippocampus, a brain structure part of the limbic system located in the medial temporal lobe. This measure is still regarded as an early hallmark predicting progression from MCI to dementia in a clinical setting (Petersen, 2011; Caillaud et al., 2019). Substantial volume loss in patients with MCI and AD has been confirmed by several cross-sectional and longitudinal studies (see e.g. Apostolova et al., 2012; Franko & Joly, 2013; Gorbach et al., 2020), and more generally, this brain structure is particularly vulnerable to the process of aging (Zheng et al., 2018), with an accelerating volume loss in the middle age (Nobis et al., 2019).

The critical role of the hippocampus in learning and memory function (Zeidman

& Maguire, 2016) is another argument for including this volume measure in the present study. Several studies have confirmed that atrophy of the hippocampal structure correlates strongly with cognitive decline (Petersen et al., 2000). For instance, a community-based study found that among individuals with aMCI, those with volumetric measurements falling at or below the 25th percentile for their age and sex had two to three times as a high risk of progressing to dementia over a 2-year follow up compared to those whose hippocampal volume were at or above the 75th percentile (Jack et al., 2010). Taken together, these studies show the importance of including measures of hippocampal volumes when predicting a trajectory from MCI to AD.

**Figure 5**

*Atrophy of a healthy individual (A) compared to an individual with Mild Cognitive Impairment (B) and Alzheimer's disease (C). As illustrated by the arrow, cell loss causes shrinkage of the hippocampus. Illustration adapted from Petersen (2011).*




### *1.1.4 Depression in MCI and AD*

Depression is one of the most common neuropsychiatric symptoms in the elderly population, with community-based studies reporting symptoms of depression in 20% of the elderly population (Lyketsos et al., 2002). It is well described as a cardinal symptom of some of the main neurodegenerative disorders (e.g. Parkinson's disease and Huntington's

disease) (APA, 2013, p. 181). Recently, there has also been an increased awareness of how depressive symptoms can be an early sign of AD and can cause as much and sometimes more distress than the cognitive symptoms (APA, 2013). The importance of assessing symptoms of depression in patients with MCI is further underscored by a recent meta-analysis of neuropsychiatric symptoms in this patient group (Martin & Velayudhan, 2020). They referred to studies showing prevalence rates up to 83% in clinic-based samples of aMCI (Rozzini et al., 2007). Symptoms of depression may therefore be the first to get medical attention in a patient with early signs of AD.

Still, the significance of this high prevalence of depression observed in MCI is currently not clear, and findings from studies examining the role of depression in relation to the risk of progressing to AD tend to be inconclusive. Although one study found that among patients with aMCI, 85% of those with depression progressed to AD, compared to only 32% of non-depressed individuals (Modrego & Ferrández, 2004), other studies show no increased risk of progressing to AD associated with depression in aMCI patients (Palmer et al., 2010). Conflicting results are probably related to the ambiguous relationship between symptoms of depression and cognition, where depression can be considered secondary or concomitant to cognitive decline (Sachs-Ericsson & Blazer, 2014). Taken together, it is important to take depression into account when predicting a trajectory from MCI to AD, but awareness should be given its close link to the cognitive characteristics of these disorders.

### *1.1.5 Cognitive Reserve and Brain Maintenance*

Even though the presence of AD pathology in most cases leads to the clinical syndrome characterizing AD, there is a significant proportion of elderly who remain cognitively normal despite having a high load of $A\beta$ plaques and tau tangles. This is well established through several studies finding amounts of pathology sufficient to fulfill the pathological criteria for AD in individuals with normal cognition (Crystal et al., 1988; Moris et

al., 1996; Neuropathology Group, 2001). The disconnect between the degree of pathology and cognition has been recognized for a long time, and there has been a great interest in understanding potential resilience factors. Several concepts have been used to describe such resilience factors against normal and pathological age-related changes, including cognitive reserve and brain maintenance.

In this context, the cognitive reserve hypothesis has been among the most studies (Arenaza-Urquijo & Vemuri, 2018). This hypothesis posits that having greater cognitive reserves may allow for more flexible strategies in solving tasks and as such provides resilience against brain pathology (Tucker & Stern, 2011). Two individuals that seem to have similar neuropathological load can thus present with very different clinical outcomes. Brain maintenance is a concept used to describe a complementary concept to cognitive reserve (Habeck et al., 2016). It was first introduced by Nyberg and colleagues (Nyberg 2012). They referred to brain maintenance as 'hardware' and described cognitive reserve as 'software', meaning that it explains functions far beyond what can be explained by brain structure. By this distinction, the trajectory from normal cognitive performance, through MCI to AD, is an example of poor brain maintenance. The trajectory is, however, modulated by several resilience factors. With both these processes being dimensional, measures of brain volume, as well as cognitive function in samples of older adults, should always be evaluated in the context of heterogeneity. Whereas numerous studies have investigated the role of cognitive reserve and brain maintenance for the observed pathology-cognition disconnect in cognitively unimpaired individuals with AD pathology, fewer studies have assessed their relative influence specifically on MCI-individuals risk of progressing to AD (Varatharajah et al., 2019). One such study used the Japanese version of the National Adult Reading Test (NART) as an index of cognitive reserves and found that MCI subjects converting to dementia had lower premorbid intelligence compared to those who reverted. This finding indicates that cognitive reserve may be an important factor to consider when trying

to identify which individuals are on a trajectory towards AD.

Taken together, the heterogeneity of cognitive function in older adults can be explained by a wide range of unknown factors. This includes biological and genetic factors, as well as the many life events and lifestyle factors influencing an individual throughout a lifetime (Wahlhovd et al., 2014; Nyberg 2019). This gave the present study on trajectories from MCI to AD arguments for applying a comprehensive data-driven framework, including analyses of feature importance, within a machine learning approach.

## 1.2 Machine Learning

Machine learning (ML) is a branch of Artificial Intelligence in which statistical methods are used by computers to find patterns in high dimensional data. It is closely related to the field of cognitive psychology, where learning can be defined as "*the combined effect of all encoding, storage, and retrieval in gradually enhancing the performance on a particular task*" (Purves et al., 2013, p. 574), and this conceptualization of learning can be extended to the context of ML. As explained by El Naqa et al. (2015, p. 4), an ML algorithm is a computational process created to complete a specific task, and it does so by learning from input data without being explicitly programmed to do this (i.e., not 'hard coded'). An ML algorithm should rather be described as 'soft coded' because the goal is that it learns from experience (input data) to increase its performance. The 'learning' part is referred to as the training of the model. The goal is to obtain a predictive model that works on new data, i.e. data not used to train the model. To avoid that the model is overfitted to the data on which it is trained, several means can be taken. To detect whether the model is overfitting the full dataset is typically split into two parts; one part for training the model (training set) and one part used for a final validation (test set) of the model's performance. If the model performs well on the test set which contains data previously unseen by the model, this indicates that the model performance can be generalized also to new data. If

the model has poor generalization ability it is often either overfitted, in which case one would use various so-called regularization techniques, or underfitted, in which case one would attempt to increase the capacity of the model by e.g. enlarging the set of parameters or switch to another, higher-capacity model. It may also happen that the training data set distribution is too dissimilar to the test set data distribution, indicating that one must be more careful when selecting the data instances that form the test set.

Broadly, there are three types of ML algorithms: i) supervised, ii) unsupervised, and iii) reinforcement learning. Supervised learning are theoretically driven top-down approaches, in which the algorithm is trained by the use of labeled data. In classification settings, each observation in the dataset is paired to one 'true' label or *class*, and the algorithm tries to classify an outcome based on selected features (input data). Because the true labels for each observation are known to the algorithm, it can validate whether the class predicted was correct or not, and adjust accordingly. For regression models, the predicted values are continuous numbers that can be compared to the "true" values using various distance measures, for example mean squared distance between the predicted values and the true values. In contrast, unsupervised learning algorithms are empirically driven bottom-up approaches, where the input data contain no such true labels (i.e. no ground truth). Thus there are no error or reward signals to base an evaluation on, so instead, the algorithm's task is to uncover meaningful patterns in the data. This can for instance be by defining clusters of observations sharing properties in the high dimensional space of multiple input features. The last type, reinforcement learning, is the most dynamic form of ML. Here, the algorithm is an agent learning from its environment to maximize reward based on the feedback it gets from its actions. For each of these three broad categories there are numerous different ML models available, and which approach is most suitable depends on the research question at hand. In the context of this thesis, a supervised classification model was deemed appropriate as we wanted to investigate how well a model could classify MCI subjects as

stable or converters (true labels). Properties of Random Forest (RF), the specific supervised classifier used in this study, will be described in the next section. ML approaches, as compared to traditional statistical methods more commonly used, have been found suitable when trying to reveal the complex interplay between a large number of predictors (Carreiro et al., 2015). Over the past decades we have witnessed a boost in the emergence of ML approaches applied to medical research, and this is true also for the research field of AD (Dallora et al., 2017). In line with this, several studies employing ML frameworks have proven such methods to be powerful tools for predicting disease trajectories of MCI patients (Battista et al., 2017; Moradi et al, 2015; Amorosa et al., 2018). Nevertheless, despite the apparent utility of such frameworks, they have mainly been applied to studies including neuroimaging and genetic data, and fewer studies have assessed cognitive, behavioral, and daily-life functional data (Battista et al., 2017). Studies investigating these aspects in relation to MCI and AD have to a greater extent relied on traditional statistical methods (Pereira et al., 2018). One plausible reason for this relative lack of ML frameworks being extended to neuropsychological data might be that the inherent high-dimensionality of both imaging and genetic data has created a more pressing demand for novel methods of analyzing such data. Further, ML is inherently a multidisciplinary field drawing on knowledge from several different domains such as statistics, computer science, and engineering, as well as domain knowledge from experts in the specific topic being studied. We therefore speculate if fewer studies employing ML frameworks on neuropsychological data may be due to greater disciplinary distance, and hence less interdisciplinary cooperation between computer scientists and clinical neuropsychologists compared to the field of imaging and genetics.

There are, however, several reasons why ML frameworks should also be extended in this context. Cognitive impairments are cardinal symptoms of both MCI and AD, and the core part of the clinical picture first meeting the clinician. In the ‘real world’ we do

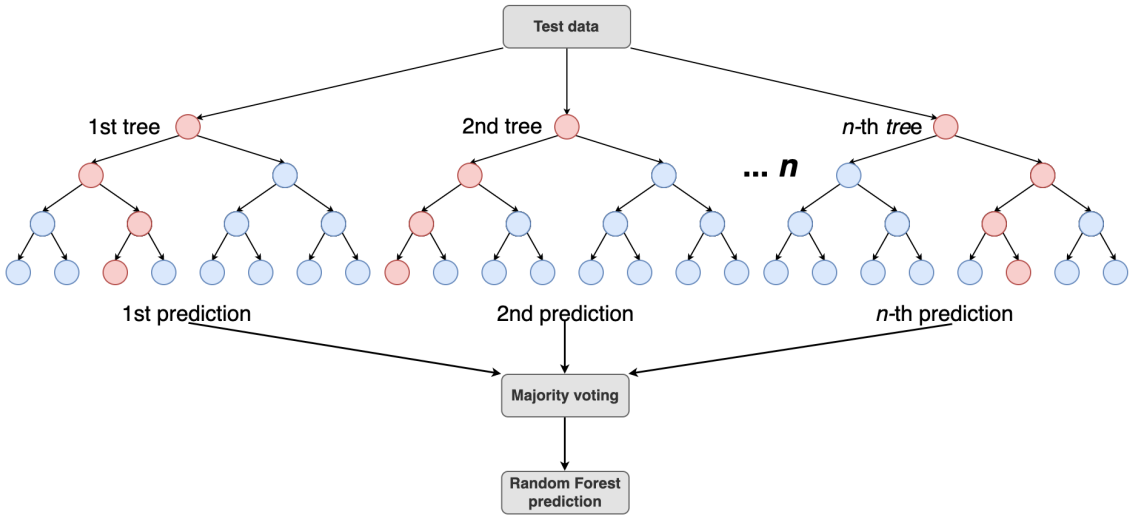
not screen the population for AD pathology, and the initial cognitive symptoms of MCI may therefore be the first chance to capture individuals at a predementia stage. Neuropsychological tests are therefore widely used in the clinical setting (compared to more expensive and/or invasive biological markers such as structural or functional MRI imaging, PET scans, or cerebrospinal fluid). Due to restricted time, clinicians may find it difficult to select which neuropsychological tests, among the countless tests available, are most important for prognostic prediction. This gives arguments for the need for data-driven approaches to identify feature importance when investigating the relationship between subject-specific information at baseline and disease trajectory.

### *1.2.1 Random Forest*

Random Forest (RF) is a commonly used supervised ML model introduced by Breiman (2001a). This is an *ensemble* model in which multiple decision trees are built, from which each tree in the ensemble casts a vote on class belonging. As illustrated in Figure 6, the final prediction of class belonging is decided based on majority voting, meaning that the predicted class of a given observation is the one that the majority of trees voted for. When constructing trees in an RF, the concept for maximizing information gain in each split is done in the same way as when creating a single decision tree. That is, the goal is to optimize information gain (i.e. decreasing impurity of the split) at each node in the tree. This is done by selecting the most informative feature, as well as the most optimal value of this, to split on. An advantage of RFs is thus that they harness' the simplicity associated with decision trees. However, they introduce some randomness, which typically results in better predictions as each predictor in the ensemble has a different decision logic. The randomness is introduced mainly in two ways. Firstly, each tree in the forest is grown based on drawing a *bootstrapped* dataset from the full training data. Creating bootstrapped samples means drawing only a random subset of observations (i.e. subjects) from the orig-

**Figure 6**

*Illustration of Random Forest.*




inal training data with replacement. An important aspect of bootstrapping is that the same observation can be selected several times, hence each of the bootstrapped samples will contain the same number of observations as the original training data, while not being identical due to duplicate entries being allowed (Hastie et al., 2009, p. 249). For each bootstrapped sample, there will also be observations not selected. These are called *Out-Of-Bag* samples and are run down the constructed tree to provide an estimated accuracy of the tree. Secondly, each tree in the ensemble only gets access to a random sample of the features in the data set.

Evaluating how well an ML algorithm performs in a classification task is an important part of the process. Several approaches and metrics for evaluating classification models exist, and which are most informative in a given case is closely related to both the research question and characteristics of data used in model construction. One central metric of model performance that will be reported in the present study is *accuracy*, namely the percentage of correctly classified subjects. However, as argued by Japkowicz

& Shah (2009), the use of additional metrics is often necessary to get a nuanced assessment of a model's strengths and weaknesses. This is for instance true in cases where data is unbalanced with respect to classes, in which accuracy can be a poor indicator of model performance. Therefore the *F1-score*, which is a harmonic mean between positive and negative predictive value, will also be reported. Additionally, *sensitivity* and *specificity*, two performance metrics central in medical classification problems, will be reported. Sensitivity gives information about the proportion of positive cases that are correctly identified (i.e. true positive rate), and conversely, specificity gives information about the proportion of negative cases that are correctly identified (i.e. true negative rate).

## 1.3 Problem Formulation and Objectives

Based on the discussion above it should be clear that MCI is a heterogeneous diagnostic construct with an uncertain course of development on an individual basis. Being able to identify individuals at increased risk for developing AD is of great importance in the context of precision medicine. Taken together, this motivated the current thesis to explore the following three research questions:

**RQ1:** Is there a group difference in the clinical phenotype of MCI subjects remaining stable (sMCI), compared to MCI subjects converting to AD (cAD), already at baseline?

**RQ2:** How well can a Random Forest machine learning algorithm trained on baseline data perform in the binary problem of classifying MCI subjects into those who will remain stable (sMCI) and those who will convert to AD (cAD)?

**RQ3:** What features are weighted highest in making this prediction?

# 2 Methods

## 2.1 ADNI database

All data for the current study was obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI), one of the world's leading projects within research on MCI and AD. This is an ongoing longitudinal study initiated in 2004 as the result of collaboration between several academic institutions and private companies. It is a non-randomized natural history study where participants do not receive any treatment but their health data is being collected longitudinally to understand the natural developmental trajectory from normal cognition to AD. One of the project's main objectives is to develop markers for early detection and monitoring of people who are on a trajectory towards AD. To achieve this, clinical, genetic, brain imaging, and biological data in the form of cerebrospinal fluid and blood samples have been collected longitudinally from participants at 59 different research centers in the United States and Canada.

Originally, ADNI was meant to last for five years (ADNI 1 from 2004-2009), but before the first study wave was completed, the project received funding to be extended for three subsequent phases: ADNI-GO (2009-2011), ADNI-2 (2011-2016) and ADNI-3 (2016-2021). To date, these four protocols have recruited over 2000 elderly with i) normal cognition, ii) early or late MCI and iii) people with early AD.

Many of the subjects originally enrolled in ADNI 1 are also followed in subsequent study waves, and new subjects have been enrolled in each of the subsequent phases. An aim of ADNI was to keep the study protocols similar across the different study phases, but certain updates and modifications have been found necessary due to improved knowledge and technological advances. This has challenged longitudinal studies exploiting data from the ADNI database because subjects with complete data from one study phase in many cases miss data points from another phase. This challenge is also valid for the present

study. Much time and effort were therefore put in the first step of data preparation, with an aim to include as many participants as possible across all four phases.

To be enrolled in the ADNI study, all subjects had to pass a screening in which the following inclusion criteria had to be met: i) age between 55-90; ii) Hachinski Ischemic Score less than or equal to 4; iii) Geriatric Depression Scale less than 6; iv) study partner with a minimum of 10 hours contact per week either in person or telephone, who also could accompany to study visits; v) visual and auditory acuity adequate for neuropsychological testing; vi) good general health with no diseases prior to enrollment; vii) women had to be sterile or two years past childbearing potential; viii) being willing and able to complete a 3 year imaging study (2 years for AD subjects); ix) having a minimum of 6 grades of education or work history equivalent to this; x) being fluent in either English or Spanish; xi) commitment to Neuroimaging and no medical contraindications to MRI; xii) agree to provide DNA for ApoE testing and banking of genetic material, as well as blood and urine for biomarkers; and xiii) not presently being enrolled in other trials or studies.

Further, all subjects had to be stable on permitted medications for at least 4 weeks prior to screening. For subjects with MCI and AD permitted medications included Cholinesterase inhibitors and Memantine. For all participants, estrogen, and estrogen-like compounds, as well as vitamin E substitutions, were permitted (see http://adni.loni.usc.edu/methods/documents for full list of permitted medications).

## 2.2 Participants included in the present study

For the present study we included subjects across all four study phases of ADNI who according to ADNI's criteria were defined as MCI at their baseline (first) assessment. Data were downloaded on November 9th 2020, and the study is thus restricted to subjects whose data was uploaded to the ADNI database before this date.

ADNI defined a subject with MCI if; i) s/he or her/his partner reported concern due

to impaired memory function; ii) s/he obtained a Mini Mental State Examination (MMSE) score between 24 and 30; iii) a Clinical Dementia Rating Scale (CDR) score = 0.5; iv) a score lower than expected (adjusted for years of education) on the Wechsler Memory Scale Logical Memory II (WMS-II); and v) had preserved function of daily living. From this group of MCI subject we selected subjects who met the additional criteria of having at least three study visits (e.g. baseline visit and at least two additional visits) and who had undergone a minimum of three MRI examinations.

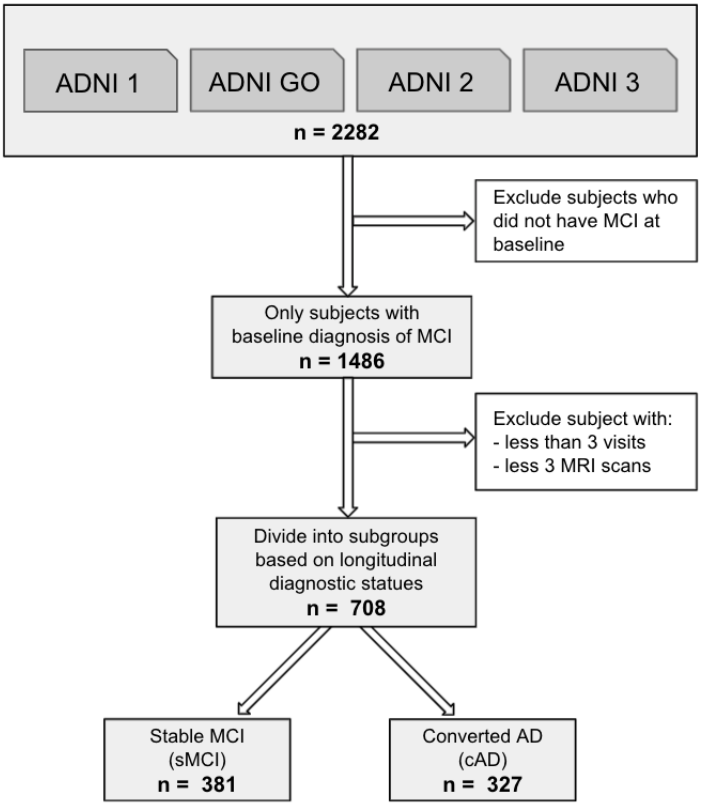
These MCI subjects were further divided into two subgroups defined according to their longitudinal diagnostic status. One subgroup was defined as stable MCI (sMCI), meaning that they met the applied ADNI criteria for MCI on all study visits (n=381, age range at baseline = 55-91). The other group was defined as converters to AD (cAD) and included subjects who initially were diagnosed with MCI, but converted to AD at a later study wave (n=327, age range at baseline = 55-88). ADNI defined AD by the following criteria: i) an MMSE score between 20-26 (inclusive), ii) a score = 0.5 or 1.0 on CDR, and iii) when they met the National Institute of Neurological and Communication Disorders and Stroke/Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria for probable AD (McKhann et al., 1984). To ensure uniform application of diagnostic criteria across the over 59 different study cites involved, a Central Review Committee verified each individual subject's conversion to AD. Figure 7 illustrates the process of subject selection and creation of subgroups.

### 2.3 Neurocognitive Measures

When selecting participants for the present study, we aimed to include as many as possible with data on validated neuropsychological tests known to be affected in patients with MCI and AD. Due to the aforementioned challenges related to differing study protocols across the four ADNI phases, there was however a trade-off between sample size

**Figure 7**

*Flowchart illustrating the process of selecting subjects and creating MCI subgroups.*




and tests to be included. In the end, the Rey Auditory Verbal Learning Test (RAVLT) was included to assess main aspects of the process of memory function: immediate recall, delayed recall and delayed recognition, and different aspects of attention/executive function were assessed by the performances on the Trail Making Test part A and B and a semantic fluency test. In addition, we included two more global measures: the short form of the Geriatric Depression Scale (GDS), and the American National Reading Test (ANART) to assess symptoms of depression and intellectual function, respectively. All included tests are commonly used clinically, and all examinations were conducted by certified personnel. Each of the selected neurocognitive tests and the individual scores derived will be described in the paragraphs below.

### *2.3.1 Rey Auditory Verbal Learning Test (RAVLT)*

RAVLT (Rey, 1964) is a list learning task included as a measure of different aspects of verbal learning and memory function. In the first learning trial, a list of 15 nouns is read aloud by the test administrator at a rate of one word per second. Immediately after the first presentation, the subject is asked to freely recall as many of these 15 words as possible. This procedure, with reading and recall of the same list, is repeated for 4 more trials. A total score for immediate recall [`'RAVLT immediate'`] was calculated by adding the number of words correctly recalled across all five trials. After a 30-minutes delay period filled with testing unrelated to the verbal content of RAVLT, the subject is again asked to recall the 15 words from the original list, and the number of correct responses is used as a measure of delayed recall [`'RAVLT delayed'`]. Immediately following this, a list including the 15 targeted words intermixed with 15 distractor words is presented to the subject who is asked to circle the words s/he recognizes. From this, a recognition score was derived from the sum of correct answers [`'RAVLT recognition'`].

### *2.3.2 Trail Making Test (TMT)*

TMT (Reitan, 1958) was included as a measure of processing speed and executive function. This assessment consists of two parts, TMT-A and TMT-B, which both depend on visuomotor and perceptual-scanning skills and tempo, but where part B adds a load on the cognitive flexibility part of the executive function.

In part A, a sheet of paper with the numbers 1-25 printed on it is presented to the subject. The subject is then instructed to use a pen to connect the numbers in ascending order, encouraged to work as fast as they can. Part B is similar, but here the numbers (1-13) are intermixed with letters (A-L), and the subject is instructed to connect these by switching between the ascending numerical and alphabetical order (i.e. 1 to A, A to 2, 2 to B). If an error is made during the test session, the examiner stops the subject and redirects

him/her back to the last correct response. The total number of seconds used to complete the tasks was given separately for part A [`'Trail Making A'`] and B [`'Trail Making B'`]. Maximum (worst) scores are 150 and 300 for part A and B, respectively, as the subject was stopped if these time limits were exceeded. In the present study the time spent to complete TMT-B is used as a measure of executive function, although we are well aware that the performance is dependent on several cognitive abilities such as processing speed, sequencing, mental flexibility, and visual-motor skills (Bowie & Harvey, 2006).

### *2.3.3 Category Fluency Test (CFT)*

CFT (Butters et al., 1987) assess verbal fluency. In CFT, the subject is asked to generate as many exemplars as possible of words belonging to a given semantic category (animals) within a testing period of 1 minute. A primary performance measure [`'Category Fluency'`] was calculated based on number the of correct, unique examples generated. The validity of CFT to assess verbal ability, and more specifically lexical access ability, has been confirmed in several studies (Lezak et al., 2012, p. 693; Shao et al., 2014). However, the task does not only tap into the domain of language but it is also heavily dependent on executive function (Baldo & Shimamura, 1998; Schwartz & Baldo 2001). This is because in addition to accessing their mental lexicons, the subjects must focus on the task at hand, select words meeting the condition of belonging to the semantic category, and inhibit repetitive responses.

### *2.3.4 Geriatric Depression Scale (GDS)*

The short form of the GDS (Yesavage & Sheikh, 1986) is a self-report questionnaire designed to identify symptoms of depression, specifically in an elderly population. The form includes 15 items to which the subjects answer by circling "yes" or "no" based on how they felt the past week. Ten questions are positively oriented for depression (e.g. "Do

you feel that your life is empty?") and the remaining five questions are negatively oriented (e.g. "Are you basically satisfied with your life?"). All questions are weighted equally, with one point given for each answer indicative of depression (maximum 15 points). As participants obtaining a total GDS score [`'GDS'`] between 6-15 were already excluded from the ADNI sample, the total GDS scores in our selected sample range between 0-5. The score in individual participants are still used to assess severity level, as even symptoms below diagnostic threshold may affect cognitive function (Brevik et al., 2013)

#### *2.3.5 American National Adult Reading Test (ANART)*

ANART (Nelson & O'Connell, 1978) estimates intellectual function by asking subjects to read a list of 50 words that are printed on a sheet of paper. All words are irregular in that they do not follow rules of phonography and orthography, and they are graded in terms of difficulty of correct pronunciation. Because of this irregularity, correct pronunciation can not be achieved by applying common grammatical rules, but rather depends on previous familiarity with the words. Performance is assessed according to phonetic accuracy in pronunciation of each word, and a total score [`'ANART'`] was calculated in terms of the total number of committed errors.

### 2.4 MRI acquisition and Brain Segmentation

Acquisition of 1.5 T MRI (for ADNI 1) and 3.0 T MRI (for ADNI GO/2/3) data at each of the multiple ADNI sites followed a described standardized protocol developed by ADNI. See `http://adni.loni.usc.edu/methods/mri-analysis/mri-acquisition` for sequence details.

To extract reliable hippocampus volume estimates, T1-weighted MRI images were automatically processed with the longitudinal stream (Reuter et al., 2012) in FreeSurfer v.7.1.1. Specifically, an unbiased within-subject template space and image

(Reuter & Fischl, 2011) is created using robust, inverse consistent registration (Reuter et al., 2010). Several processing steps, such as skull stripping, Talairach transforms, atlas registration as well as spherical surface maps and parcellations are then initialized with common information from the within-subject template, significantly increasing reliability and statistical power (Reuter et al., 2012). ADNI data were originally processed with two different versions of FreeSurfer (v.4.3 and v.4.1). As shown in previous work from colleagues (Mofrad et al., 2021), the use of various versions of FresSurfer may lead to larger discrepancy in the atrophy estimations. Thus, all included MRI images were re-processed applying the same version of FreeSurfer (v.7.1.1) by collaborators at the Mohn Medical Imaging and Visualization Centre. A measure of total the hippocampus volume [`'Hippocampus'`] was derived by combining the volume of the left and right hippocampi. To reduce the effect of individual and gender differences in brain sizes, the volumes were normalized using a total intracranial volume measure estimated (eTIV) by Freesurfer.

### 2.5 ApoE Status

Blood samples were collected at baseline for ApoE genotyping. Samples were transported from each study site by overnight transport to the University of Pennsylvania Alzheimer's Disease Biomarker Laboratory where the genotyping was carried out. In the present study, ApoE-$\varepsilon$4 status was divided into a binary variable [`'ApoE 4'`] of subjects having no $\varepsilon$4 alleles (ApoE negative) and subjects having at least one $\varepsilon$4 allele (ApoE positive).

### 2.6 Analytic Approach

The exploratory statistical analysis was performed using IBM SPSS Statistics for Macintosh, Version 27.0. The supervised data-driven machine learning analysis was implemented in Jupyter Notebooks using Python (3.5.4), Numpy (1.20.1), Pandas (1.2.4),

Statsmodels (0.8), Scikit-learn (0.19), Scipy (1.6.2), Seaborn (0.11) and Eli5 (0.11.0). The packages Matplotlib (3.3.4) and Pdpbox (0.2.1) were applied for producing figures. Relevant Jupyter Notebooks are available on the project's GitHub repository (`https://github.com/ingryy/mci_subgrups.git`).

### *2.6.1 Explorative data analysis*

A core objective of the current study was to provide a broad phenotypic characterization of the two MCI subgroups (i.e. the cAD and sMCI groups) at baseline, and compare the groups on these characteristics. The groups were therefore checked for similarities and differences with respect to all demographic and clinical measures. Student's t test for independent samples was used for continuous variables, and Pearson Chi-Square test for nominal variables. Statistical analysis of the fourteen included variables were Bonferroni corrected for multiple comparisons, with an alpha level of .004 ($\alpha_{altered} = .05/14 = .004$, rounded) considered to be statistically significant. To check pairwise correlations between the cognitive measures, Pearson correlations were calculated and presented separately for the sMCI and cAD groups in a comprehensive generalized pairs plot.

### *2.6.2 Prediction of MCI subgroups*

Prior to constructing the RF model, we found the quantity of missing data to be less than 5% and used descriptive statistics to identify potential distributional outliers. In total, 30 subjects had missing values on one or more features included in the RF model, and these were removed from the dataset prior to model construction. This resulted in a sample of 678 subjects.

We used a Random Forest (RF) classifier as implemented in `Scikit-learn` to predict class $y_i \in \{\text{sMCI, cAD}\}$ from a feature vector $\mathbf{x}_i = (\text{Age}_i, \text{gender}_i, \text{RAVLT-immediate}_i, \text{RAVLT-delayed}_i, \text{RAVLT-recognition}_i, \text{Trail-Making-A}_i, \text{Trail-Making-B}_i, \text{Category-}$

Fluency$_i$, GDS$_i$, ANART$_i$, Apoe-4$_i$, Hippocampus-volume$_i$) where $i \in \{1, ..., 678\}$ denote participant number $i$. A detailed description of the specific classifier used is found in the `Scikit-learn` Package's own documentation: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html` and the references therein. Because the dataset was relatively well balanced with respect to percentage of subjects belonging to each class (sMCI 53.7%; 47.3% cAD), the accuracy metric was used to assess model performance during development and selection of hyperparameters (explained below).

It is well known that learning the parameters of a classification function and testing it on the same data is a methodological mistake (Lundervold & Lundervold, 2019). Such a model would learn (i.e. memorize) the labels of the sample it was trained on, leading to a perfect score on this data, while potentially failing to predict anything useful when tested on unseen data. This is, as previously explained, known as overfitting, and can lead to a lack of generalization abilities. To avoid this, we split the complete sample ($n = 678$) into a training set comprising 80% ($n = 539$) used for training the mode, while a test set comprising 20% ($n = 139$) was held aside to be used for a final evaluation. This was done to assess how well the model performs on unseen data. The training and test sets were carefully stratified with respect to age, gender and class belonging. Exploratory analysis revealed no significant differences on any of the features included, nor length of follow-up.

### *2.6.3 Tuning model hyperparameters using grid search*

The RF algorithm has several hyperparameters that can be adjusted in order to optimize the classifier. Therefore, to improve model performance, we conducted tuning of the algorithm's hyperparameters. This was done by utilizing the `GridSearchCV` available from `Scikit-learn` (`http://scikit-learn.org`). Through this method, all possible combinations of the parameter values within a defined space to search are evaluated to identify

**Table 1**

*The table presents the defined range on which grid search was conducted for each of the hyperparameters. The rightmost column presents the optimized values for each parameter.*

| Parameter | Defined search range | Selected values |
|---|---|---|
| n_estimators | [100, 200, 300, 450, 470, 480, 490, 500, 550, 800, 1000] | 480 |
| max_features | [1, 2, 3, 4, 5] | 4 |
| min_samples_split | [1, 2, 3, 4, 5] | 2 |
| max_depth | [1, 2, 3, 4, 5] | 4 |
| min_samples_leaf | [1, 2, 3, 4, 5] | 1 |
| bootstrapping | [True, False] | True |

which combination of hyperparameter values results in the greatest model accuracy. In the current study, exhaustive grid searches were conducted on the following six parameters: i) number of trees in the forest (`n_estimators`); ii) number of features to consider at each split (`max_features`); iii) maximum number of levels in each tree (`max_depth`); iv) minimum number of subjects placed in a node before it can be split (`min_samples_split`); v) minimum number of subjects allowed in a (`min_samples_leaf`); and vi) whether bootstrapping should be employed (`bootstrap`). Before grid search was conducted, the accuracy we obtained was 65.1% and this increased to 73.3% after implementing optimal values for all six parameters. Table 1 presents the search space defined as well as the selected values for each of the six parameters.

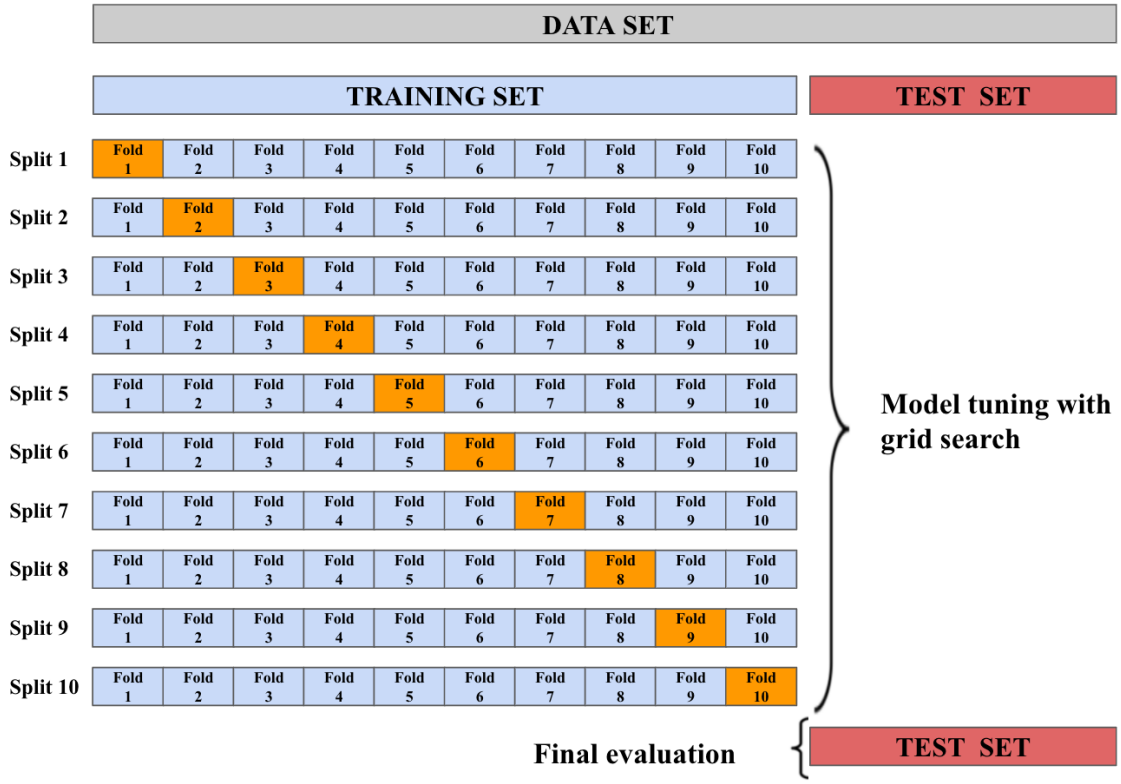
### *2.6.4 Evaluation using K-fold cross validation*

Evaluation of different parameter settings for optimizing a model can provide biased performance measures during the grid search, as the performance has to be checked against a held-out validation data set that may not be a good representation of the real data distribution. Each parameter setting in the grid search was therefore evaluated multiple times on different subsets of the training data set using *K*-fold cross validation.

This was achieved by dividing the training set into *K* equally sized folds, from which data from *K*-1 folds were used for training the algorithm, with the remaining *K*th fold being used for validation (see Figure 8). In this study, we stratified the training set into ten folds (*K*=10) by preserving the same ratio of the two classes in each fold.

**Figure 8**

*Illustration of the K-fold cross validation algorithm.*




### *2.6.5 Feature importance*

After establishing how well the RF model performs on classifying the two sub-groups of MCI, we further assessed the prediction importance of the 12 features included in the model. Tree-based models for feature importance, including RFs, investigate to which degree each feature decreases impurity at a splitting node. However, as pointed out

by Strobe et al. (2007; 2008), this method tends to artificially inflate the importance of features if predictor variables vary in measurement scales and/or number of categories.

We therefore added permutation testing when assessing feature importance, a technique introduced by Breiman (2001a; 2001b). Here, the effect of each feature on model accuracy is quantified by randomly reshuffling each predictor variable (one at the time), while assessing how this affects model performance. As random shuffling breaks the true relationship between a given feature and the outcome, model accuracy will decrease when a feature with true predictive power is permuted, whereas permuting a non-informative feature will likely render model performance unchanged, or even improved.

# 3 Results

## 3.1 Exploratory Analysis

A total of 708 subjects defined as MCI at baseline met our inclusion criteria and were included in the current study. From this sample, 381 were labeled sMCI and 327 were labeled cAD based on whether they remained stable at MCI or converted to AD during their participation in ADNI.

### *3.1.1 Demographic characteristics by subgroups*

As shown in Table 2, there was no statistically significant difference in age between the two groups, with both groups having a mean of approximately 73 years at baseline. Further, mean education was almost 16 years and the percentage of females approximately 40% in both groups. None of these differences in demographics were statistically significant. Subjects defined as sMCI had a significantly shorter follow-up time than the cAD group.

**Table 2**

*Descriptive and Comparative Statistics for demographic and clinical data. Analysis conducted with Student's t test for for independent samples ($\chi^2$ test for the categorical variables Gender and Apoe 4).*

| | sMCI (n=381) | cAD (n=327) | | | |
|---|---|---|---|---|---|
| | *Mean ± SD* | *Mean ± SD* | *t* (df) / $\chi^2$ (df) | *p* value | *Effect size* |
| **Demographics** | | | | | |
| Age | 73.0 ± 7.51 | 73.9 ± 7.07 | 1.59 (706) | .113 | 0.12 |
| Education | 15.90 ± 2.90 | 15.90 ± 2.76 | 0.0461 (706) | .963 | < 0.01 |
| Female, % | 40.4 | 39.4 | 0.0691 (1) | .793 | 0.01 |
| Length of follow-up, y | 4.53 ± 2.72 | 5.07 ± 2.72 | 2.58 (706) | .010 | 0.20 |
| **Memory** | | | | | |
| RAVLT immediate | 36.7 ± 10.6 | 29.3 ± 7.74 | 10.4 (706) | < .001 | 0.78 |
| RAVLT delayed | 4.88 ± 3.90 | 2.02 ± 2.66 | 11.2 (706) | < .001 | 0.85 |
| RAVLT recognition | 11.3 ± 3.15 | 9.39 ± 3.57 | 7.47 (705) | < .001 | 0.56 |
| **Executive function** | | | | | |
| Trail Making A | 39.4 ± 15.6 | 44.6 ± 21.3 | 3.78 (706) | < .001 | 0.29 |
| Trail Making B | 109 ± 58.6 | 133 ± 73.7 | 4.74 (696) | < .001 | 0.36 |
| Category Fluency | 17.8 ± 5.19 | 15.8 ± 4.73 | 5.12 (706) | < .001 | 0.39 |
| **Global measures** | | | | | |
| GDS | 1.73 ± 1.45 | 1.66 ± 1.39 | 0.698 (706) | .486 | 0.05 |
| ANART | 13.10 ± 9.48 | 13.20 ± 9.62 | 0.244 (700) | .807 | 0.02 |
| **Biomarkers** | | | | | |
| Apoe 4 postive, % | 42.3 | 64.2 | 34.0 (1) | < .001 | 0.22 |
| Hippocampus | $0.00452 \pm 7.56{*}10^{-4}$ | $0.00398 \pm 6.80{*}10^{-4}$ | 9.81 (692) | < .001 | 0.75 |

*Note.* Effect sizes are reported as Cohen's d for continuous variables and Cramer's φ for categorical variables. Abbreviations: sMCI = stable Mild Cognitive Impairment; cAD = converted to Alzheimer's disease; RAVLT = Rey Auditory Verbal Learning Test; GDS = Geriatric Depression Scale; ANART = National Adult Reading Test.

### *3.1.2 Global measures*

The group means for American National Reading Scale (ANART) and Geriatric Depression Scale (GDS) are shown in Table 2. Group comparisons showed no statistically significant differences on any of the two measures.

### *3.1.3 Memory function and attention/executive function*

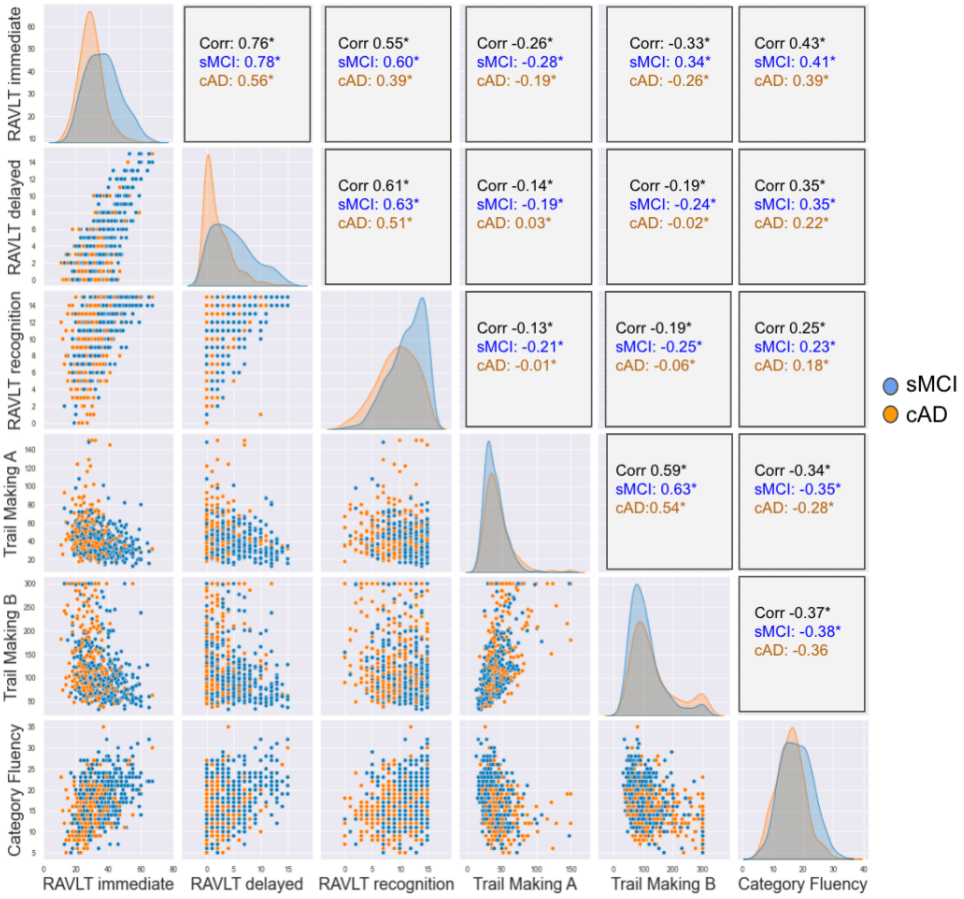
The means in performance on measures of memory function and executive function are shown in Table 2. Compared to the cAD group, the sMCI group showed higher performance on all the selected neurocognitive tests. All these differences were statistically significant at an alpha-level less than 0.001. For all three memory tests the associated effect

sizes were medium to large, whereas the effect sizes were weaker for the tests of executive function.

Figure 9 illustrates the distributions and pairwise correlations between all included scores on the neurocognitive tests, marked in separate colors for the cAD and sMCI groups. As expected, the strongest correlations were found between the measures of memory function from the RAVLT, and between the A and B version of the Trail Making Test. The cAD group showed weaker correlations compared to the sMCI group for all measures. All correlations were statistically significant at a Bonferroni adjusted alpha level ($\alpha_{adjusted} = .008$).

**Figure 9**

*Pairs plot illustrating the pairwise relationship between the six measures of cognitive function plotted by subgroup. Statistical significance at a Bonferroni adjusted alpha level is indicated by asterisks.*



### *3.1.4 Biomarkers*

A chi-square test of independence revealed that the presence of ApoE-ε4 alleles differed significantly across the two groups, with 64.2% of cAD subjects being carriers of at least one ε4 allele, compared to 42.3% of the sMCI group. The volume of the hippocampus also varied across the two subgroups, with cAD subjects on average having significantly smaller volume compared to the sMCI. This difference was accompanied by a large effect size, suggesting a high practical significance for this latter difference.

## 3.2 Random Forest Prediction Model

Results of the model performance from the cross validation procedure and evaluation on the test set are presented in Table 3. Figure 10 illustrates the accuracy of our RF classifier on the test set (red) and mean of the 10 cross validation sets (green) compared to three different 'dummy-models' (blue). The figure illustrates that the model we constructed had better accuracy than all three null-models.

**Table 3**

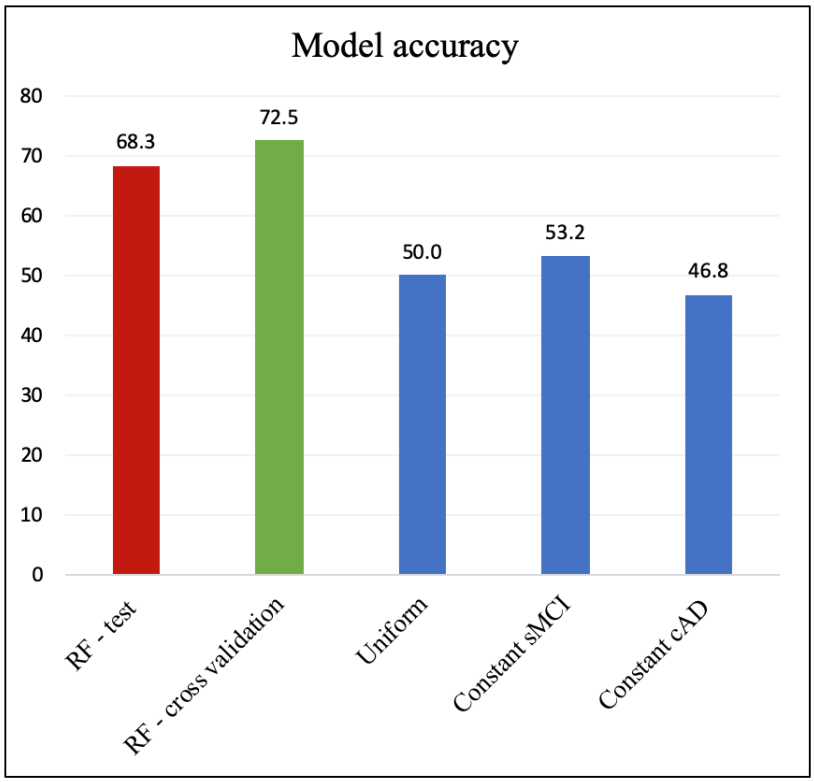
*Table depicting features included in the RF classifier model and performance metrics from cross validation procedure and the test set. Scores from cross validation are reported as mean and standard deviation across the ten folds.*

| FEATURES INCLUDED | | | | MODEL PERFORMANCE | |
|---|---|---|---|---|---|
| **Cognitive** | **Demographic** | **Global** | **Biomarkers** | **Cross validation, M (SD)** | **Test set** |
| Memory | Age | GDS | Hippocampus | Accuracy: 0.725 (0.86) | Accuracy: 0.683 |
| RAVLT immediate | Gender | ANART | ApoE 4 | F1: 0.721 (0.96) | F1: 0.667 |
| RAVLT delayed | | | | Sensitivity: 0.758 | Sensitivity: 0.677 |
| RAVLT recognition | | | | Specificity: 0.696 | Specificity: 0.657 |
| Executive Function | | | | | |
| Trail Making A | | | | | |
| Trail Making B | | | | | |
| Category Fluency | | | | | |

When the model was applied to the test set, we achieved an overall classification

**Figure 10**

*Bar plot illustrating accuracy score in percentage for the Random Forest predictor on the test set (red) and the mean of the cross validation procedure (green), compared to three dummy predictors (blue). "Uniform" predicts with equal probability that a subject will/will not convert, the "Constant sMCI" always predicts sMCI (majority class), and conversely the "Constant cAD" always predicts cAD (minority class).*



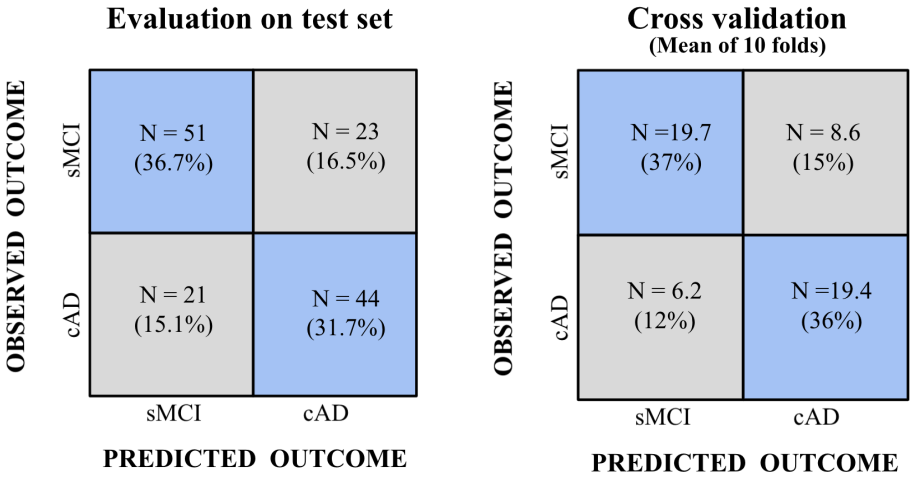

accuracy of 68.3%. The $2 \times 2$ confusion matrix (Figure 11) is computed to illustrate the true labels versus the labels returned from the RF model. The model correctly classified $\frac{44}{65}$ of cAD subjects and $\frac{51}{74}$ of sMCI subjects, resulting in a sensitivity of 67.7% and a specificity of 68.9%. The model misclassified 23 stable subjects as converters and 21 converters as stable, resulting in an F1-score (harmonic mean) of 66.7%.

Figure 13 and 12 illustrate the ranking of feature importance with respect to their importance in predicting class belonging from the cross validation (mean across 10 folds) and test procedures, respectively. In both plots, we find that the four features ranked highest

include hippocampus volume, the immediate and delayed recall subtests from RAVLT, and the B-part of the Trail Making Test.

**Figure 11**

*The blue cells represent correctly classified subjects and the grey cells represent misclassifications. The number of occurrences in each cell is given as number of subjects and percentage of the total data set for test and train, respectively.*



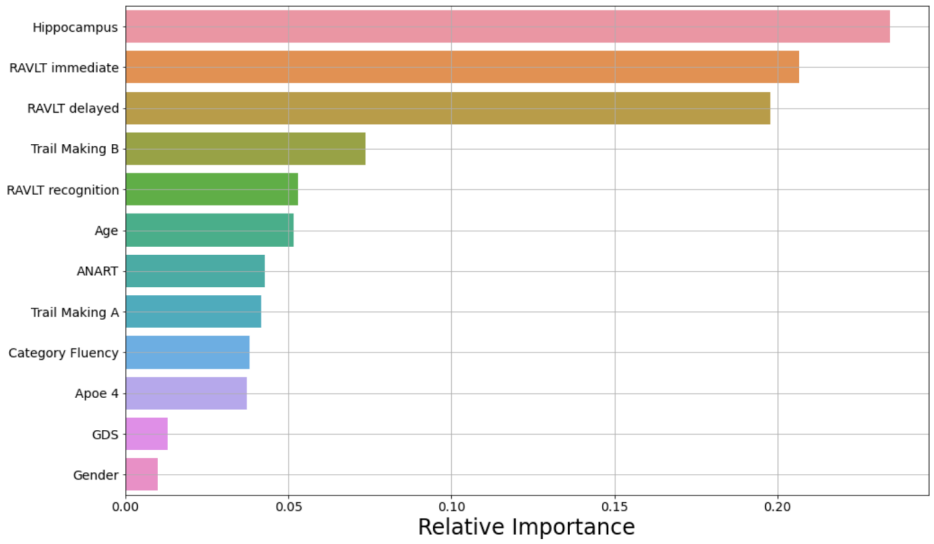

(a) *2×2 confusion matrix computed for the sMCI and cAD labels returned from prediction on test set compared with the co-occurrences of the observed outcome.*

(b) *2×2 confusion matrix computed for the sMCI and cAD labels returned from prediction on the 10-fold cross validation compared with the co-occurrences of the observed outcome.*

A model agnostic permutation importance test was added to check the results by using an algorithm where each feature is shuffled many times, with different permutations, while all the other features are kept constant. Figure 14 shows the output of this test, with positive values meaning poorer predictions on shuffled data compared to real data, indicating the feature to be more important. Here, like in the feature importance calculated based on information gain, 'hippocampus', 'RAVLT immediate' and 'Trail Making B' are ranked among the top four. In contrast, permutation testing indicates 'RAVLT delayed' to be ranked with low importance, assumed to be due to its strong correlation with the RAVLT immediate score, making the latter explain the contribution of both features.

**Figure 12**

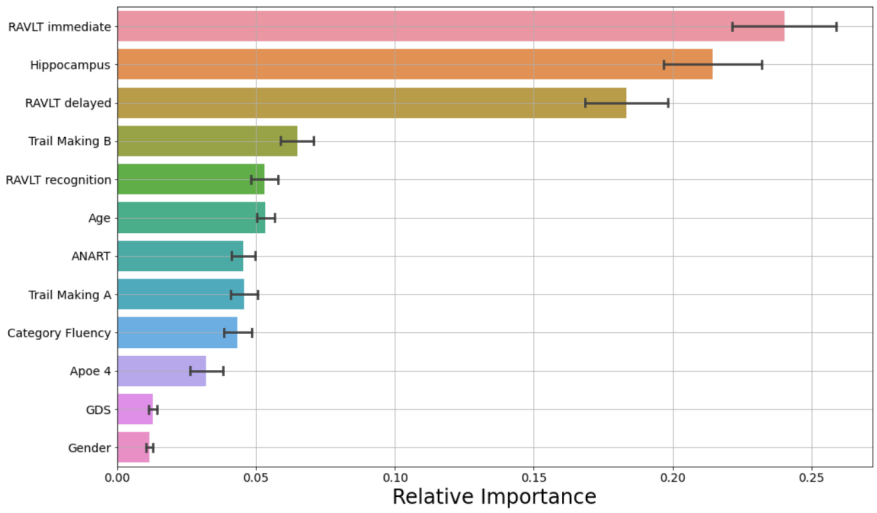
*Feature importance calculated by decrease in impurity from evaluation on test set. The 12 features included are displayed on the y-axis while the x-axis shows their relative importance.*




Partial dependence plots (PDP) were created for the three features ranked as highly important by both methods for calculating importance. These plots show the marginal effect of single features on the predicted outcome of cAD in the RF model. As such, PDPs capture the relationship between a single feature and the outcome (cAD), and are therefore informative in terms of assessing the nature of the relationship (e.g. linear or more complex). The PDP for RAVLT immediate (Figure 15) indicates that in general, remembering more words lowers the risk of converting to AD. There is a steep decline from 33 to 40 words, after which the graph stabilizes. This indicates that remembering more than 40 does not have a large impact on conversion. From the PDP of the hippocampus volume (eTIV-normalized) (Figure 16), we see that larger volume tends to decrease the risk of converting to AD. A strong trend for this is observed up to about 0.005 $mm^3$, after which the graph stabilizes, indicating that increased volume after this value does not further lower the risk of conversion. The graph in PDP of Trail Making B (Figure 17) indicates a

**Figure 13**

*Feature importance calculated by decrease in impurity reported as mean of the cross validation procedure. The black error bars indicate a 95% confidence interval. The 12 features included are displayed on the y-axis while the x-axis shows their relative importance.*




roughly linear relationship indicating that using more time on the test increases probability of converting. The steepest decline is observed from 0-75 seconds.

**Figure 14**

*The figure shows feature importance calculated by permutation. The features ranked as most important are at the top, whereas those at the bottom were ranked as less important. 'Weight' shows the average effect (and standard deviation) on model accuracy from the random shuffling. The hippocampus volume, followed by RAVLT immediate and Trail Making B were rated highest. RAVLT delayed and GDS have negative values, indicating that predictions from the shuffled data were more accurate than predictions from real data.*

| Weight | Feature |
|---|---|
| 0.0604 ± 0.0558 | Hippocampus |
| 0.0504 ± 0.0482 | RAVLT immediate |
| 0.0259 ± 0.0115 | Trail Making B |
| 0.0201 ± 0.0141 | Trail Making A |
| 0.0158 ± 0.0168 | RAVLT recognition |
| 0.0129 ± 0.0141 | Gender |
| 0.0101 ± 0.0115 | Category Fluency |
| 0.0072 ± 0.0129 | ANART |
| 0.0072 ± 0.0091 | Apoe 4 |
| 0.0043 ± 0.0147 | Age |
| -0.0000 ± 0.0257 | RAVLT delayed |
| -0.0029 ± 0.0070 | GDS |

**Figure 15**

*PDP of RAVLT immediate. The plot indicates that remembering more words lowers the risk of converting to AD. There is a steep decline from 33 to 40 words, after which the graph stabilizes at 40 words. This indicates that remembering more than 40 words does not have a large effect on conversion.*

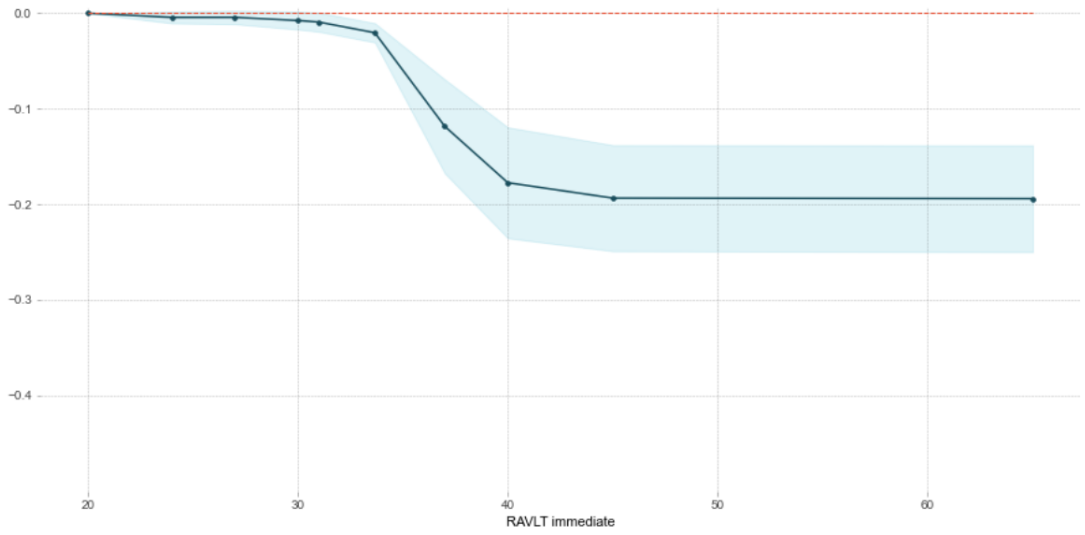


**Figure 16**

*PDP of hippocampus volume (eTIV-normalized). The plot shows that larger volume tends to decrease the risk of converting to AD up until 0.005 mm3. After this value, the graph stabilizes indicating that increased volume does not further lower the risk of progression.*

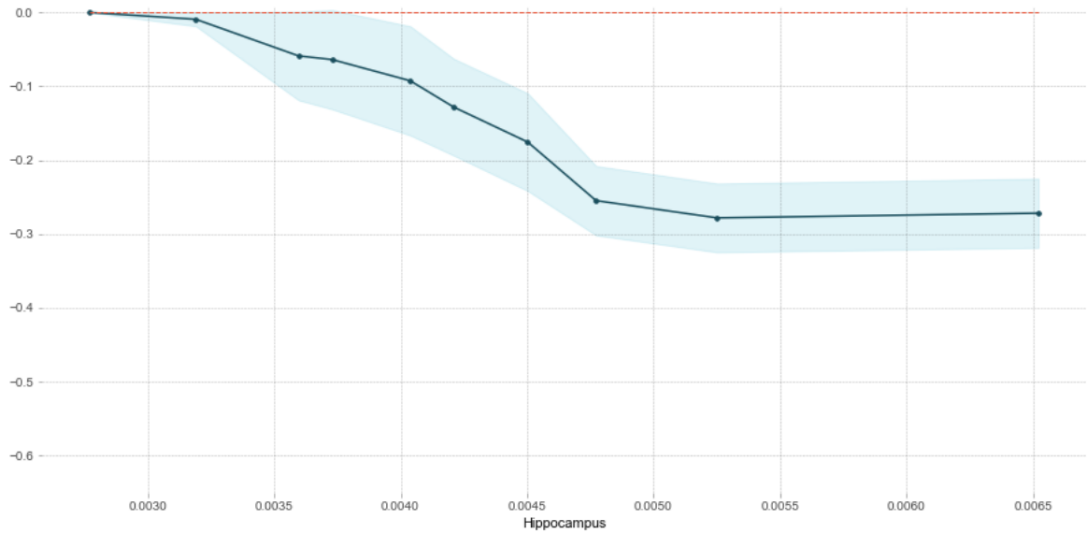
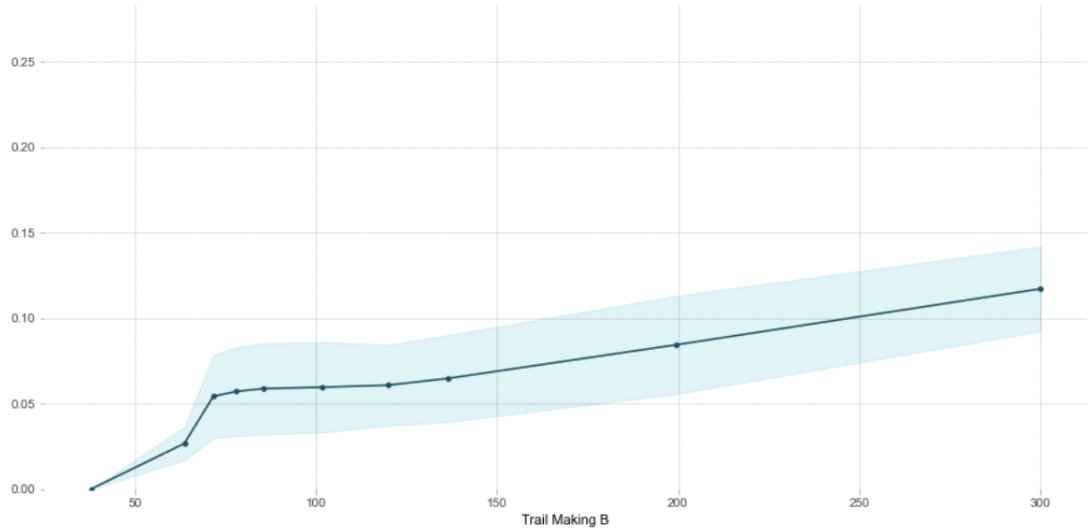



**Figure 17**

*PDP of the Trail Making B. In general, the graph indicates a roughly linear relationship indicating that using more time on the test increases probability of converting. The steepest decline is observed from 0-75 seconds.*



# 4 Discussion

The present study drew attention to two key challenges in clinical research on cognitive aging: identifying early signs of a neurodegenerative disorder and predict trajectories for individual patients showing such signs. Clinical phenotypes from an early assessment of two subgroups of aMCI patients were explored; those remaining stable (sMCI) versus those converting to AD (cAD) over time. Group differences in cognitive function could be identified already at baseline, with the sMCI group consistently performing better on included tests, compared to the cAD group. Furthermore, the cAD group showed smaller total hippocampal volume and higher frequency of ApoE positive subjects than the sMCI group, while the two groups were similar on global measures of depressive symptoms and intellectual function. We then asked how well a Random Forest machine learning algorithm trained on baseline data would perform in the binary problem of classifying individuals into these two aMCI groups, and estimated the weights of the features included in the predictive model. The group belongings were predicted with an accuracy much better than chance level (68.3%). The heaviest weights were given to features measuring immediate and delayed memory function, the total hippocampus volume, and performance on a test included as a measure of executive function, the TMT-B.

The neuropsychological data illustrated a pattern of deficits coherent with the early clinical presentation of AD described in previous studies. Even though the cAD group performed significantly worse than the stable group on all cognitive tests at baseline, our results showed that the memory tests tapping more directly into episodic memory (e.g. RAVLT immediate and RAVLT delayed) were considered most important for predicting future conversion. Although the recognition part of RAVLT relies on episodic memory, this measure was not selected among the features with the strongest weight. One hypothesis is related to the facilitating effect of familiarity in a recognition procedure, but there is

still a controversy regarding the pattern of impairments related to this memory process in MCI and AD (see for instance Yonelinas et al., 2010). In the current study, a score for recognition memory was derived from the number of targets identified. However, some studies have suggested differing neuropsychological correlates for remembering 'targets' versus 'false alarms' (i.e. identifying distractor words as targets), with the primer relying more heavily on medial temporal lobe functioning, and the latter more on frontal lobe functioning (McCabe et al., 2009). Based on this, it would be interesting for future studies to look more specifically into this difference by including an additional recognition measure for 'false alarm'. This could potentially give new insights into whether recognition memory show differential patterns of impairments in MCI subjects remaining stable versus those converting to AD. Further investigation of this could be regarded as especially relevant when considering that the other subtests of RAVLT are important predictors of conversion, and hence make this test a good candidate for a clinical tool assessing memory function. As such, the additional information gained from administering the RAVLT test could be considered a bonus in that it would not demand much extra effort in the clinical assessment.

Part B of the TMT was selected as a strong feature after those assessing memory function. This test was used as a measure of executive function (EF) in the present study. Executive dysfunction is primarily associated with an early stage of non-amnesic MCI. Over the last years, however, there has been an increased awareness that this dysfunction may be present in an early stage of AD (Guarino et al., 2019). The present study thus gives arguments for including EF tests as part of a baseline clinical examination of patients with aMCI. The importance of including a test like TMT-B is also substantiated by its dependence on other abilities known to be affected in an early stage of a neurodegenerative disease, like processing speed, sequencing, mental flexibility and visual-motor skills (Bowie & Harvey, 2006). More detailed studies should therefore map out what constitutes the more salient aspects of this and other EF tests (see e.g. Adólfsdóttir et al., 2017). As

expected from previous literature, the group of converters performed worse than the sMCI group on the CFT. As well as reflecting EF, this test relies heavily on the integrity of semantic memory. Thus, the finding that this feature was not ranked among the most important features could potentially reflect that, as opposed to episodic memory, semantic memory relies more on anterior parts of the temporal lobe, which to a larger degree is spared until the later stages of AD (Galton et al., 2001; Braak & Braak, 1991). As such, the CFT may not be among the most sensitive predictors in the early stages of MCI, but may still be informative later in the disease process, and in studies aiming to identify characteristics of different neurodegenerative disorders.

As immediate and delayed recall are measures of short- and long-term episodic memory, respectively, performance on these tests relies heavily on the integrity of medial temporal lobe structures, such as the hippocampus. This was supported by the present study, showing that all algorithms used to rank feature importance identified the hippocampus volume to be among the most important features for predicting future conversion to AD. This finding is not unexpected, as a reduction in hippocampus volume may manifest several decades prior to symptoms of impaired memory function. At the time an aMCI diagnosis is established, alertness should therefore be given to changes in memory function as well as hippocampus volume, as they probably together give important information about the risk of progressing towards AD (Mofrad et al., 2021.

In the present study we included a total volume measure of the left and right hippocampus combined. Hippocampus does, however, consist of several interconnected, but functionally and structurally distinct subfields. Recently, several studies have demonstrated atrophy in specific hippocampal subregions to be more sensitive predictors of conversion to AD. For instance, one study found that combined subfield volumes and presubiculum volume were more accurate than total hippocampal volume (Khan et al., 2014). Another study showed that the subiculum and presubiculum together have higher specificity than

the whole hippocampus in distinguishing MCI subjects who remained stable from those who converted to AD (Vasta et al. 2016). However, because manual segmentation of hippocampal substructures is a laborious and time-consuming task, the inclusion of substructure specific atrophy has predominantly been restricted to research settings. With the recent development and proven validity (see e.g. Brown et al., 2020) of fully automated techniques for hippocampus segmentation (e.g. FreeSurfer), such biomarkers could potentially be applied in clinical settings in the not too distant future.

In line with previous research, we found a significantly higher number of ApoE-$\varepsilon$4 positive subjects among those who converted, compared to those remaining stable. Despite this, ApoE status was not among the most important features for the RF model's prediction of class belonging. There are several plausible explanations for this finding. Unlike the monogenic etiology of Mendelian diseases (including some forms of early-onset AD), complex diseases such as late-onset AD, are influenced by multiple genetic and environmental factors. Heritability for late-onset AD is estimated to be up to 80% (Berkowitz et al., 2018), of which the $\varepsilon$4 of ApoE is estimated to account for 27.3% (Cauwenberghe et al., 2015). Thus, although ApoE-$\varepsilon$4 is associated with an increased risk of conversion from MCI to AD, numerous other risk genes have also been identified (Sims et al., 2020). To make genetic predictions, one should take polygenic risk scores (PRS) into account, i.e. the additive effect of a large number of genetic variants that each may have a weak effect on the phenotype of an individual (Chasioti et al., 2019). Today, several studies with sufficiently powered genome-wide datasets are conducted to shed light on the complex polygenic interplay, for instance through studies investigating how non-ApoE PRS interacts with ApoE status. One recent study found that PRS modified age for AD onset in individuals with the $\varepsilon$4, but not among non-carriers, indicating an especially detrimental effect of non-ApoE risk factors in younger ApoE positive individuals (Fulton-Howard et al., 2021). Another study found that $\varepsilon$3/$\varepsilon$3 carriers (i.e. individuals with low ApoeE-related risk) who were

classified as having the highest PRS could progress to AD a decade earlier than $\varepsilon 3/\varepsilon 3$ carriers classified to have the lowest PRS (Desikan et al., 2017). Even though ApoE is the major genetic risk factor, these studies highlight the complex and multifactorial nature of AD. Future studies are thus needed to fully untangle this complexity of the various mechanisms underlying different stages and potential subtypes of AD and its prodromal phase.

None of the two global measures of depressive symptom and premorbid intelligence, as indexed by ANART, appeared to be of importance in discriminating between the stable MCI group and the converters to AD. For the depressive symptoms, this could be related to the exclusion of participants with symptoms indicating a diagnosis of depression, but it has been shown that even less severe symptoms of depression are clinically relevant in older adults (Brevik et al. 2013). As previously discussed, the concept of 'reserve' in neuroscience of aging holds that aspects of brain structure and function can modulate the effect of neuropathology, so that individuals with greater reserves requires more severe pathology to reach the clinical threshold for cognitive impairment (Nyberg et al., 2012; Nyberg & Pudas, 2019). Knowing that both depression and intellectual function can have a mediating effect on cognitive aging, it is important to note that there is no univocal relationship between brain damage and cognitive impairment at an individual level. One can for instance hypothesize that some of the subjects that in the current study were misclassified as stable, were so due to high levels of cognitive reserves compensating for impairments. Or conversely that some of the subjects misclassified as converters showed impaired cognitive function due to depressive symptoms. It is thus possible that these two features can give valuable information about disease trajectories, but that the pattern was too unsystematic for the current machine learning framework to pick up on. Future studies should indeed look more closely into the characteristics of misclassified subjects.

The rationale and motivation behind the selection of features in the present study was closely related to the aim of keeping it clinically relevant and as a proxy of the initial

clinical assessment a person with memory concern would receive. Firstly, we restricted features to twelve theoretically motivated variables known to be involved in the process from normal cognitive aging, through MCI to a neurodegenerative disease. Secondly, we aimed to create an algorithm that potentially could be used as a tool aiding clinicians in identifying individuals at elevated risk for progression and increase knowledge about expected developmental trajectory. We are aware that including other features available in the ADNI database, such as more direct measures of amyloid load or subfields of hippocampus, could have improved the model's predictive power. Nevertheless, as this information is time-consuming, expensive, and/or invasive to obtain, it is typically not available to health professionals making diagnostic and prognostic decisions. The inclusion of such features would hence increase predictive accuracy at the expense of clinical relevance. We could also have included a broader range of neurocognitive tests. However, with time commonly being a limited resource in a "real-world" assessment at the doctor's office or in a memory clinic, we decided to restrict the inclusion of cognitive assessments to relatively brief sample of tests commonly used in a clinical setting.

The main contribution of the present study is due to the inclusion of predictive classification models. Here we showed that the RF model performed slightly better on the cross validation procedure compared to the evaluation on the unseen test set. The discrepancy of 4.2% should nevertheless be considered relatively small, which strengthens the importance of our study by lending us confidence that the model's performance was not due to being overfitted to the data it was trained on, but rather that it generalizes to new, unseen data. Further, by bringing the data into the current predictive machine learning framework, the results we obtained could be applied at a single case level and could as such be regarded as an important contribution to the clinical field of AD. Moreover, the finding that feature importances were rated quite similar by different computational algorithms gives us additional confidence in the result's validity. We will argue that this is also true

for the permutation test. Here, the immediate memory test was given a much stronger weight than the delayed subtest. We assume that this is explained by the strong correlation between the two test measures, leaving only a small contribution from the delayed subtest when the immediate memory function was selected as the primary feature. Taken together, the feature importance results underscore the importance of using methods with different algorithms when inspecting and interpreting data for feature importance.

Despite the apparent clinical utility of being able to differentiate MCI converters from non-converters, it is important to be mindful of ethical issues. Dubois and colleagues (2009) make the argument that the AD-label should encompass the full spectrum of clinical expression, including the predementia stages. They further claim that there is "*no reason to wait until the patients reach the threshold of full blown dementia for making the diagnosis of Alzheimer's disease*" (p. 136). It is however important to note that even for patients with aMCI, the outcome is uncertain and many will never develop AD. This was indeed illustrated by the confusion matrix (Figure 11a) in the present study, showing that around 15% of the participants observed as stable MCI were misclassified as converters to AD. It is important to consider that receiving a diagnosis of MCI may represent an unnecessary psychological stressor. If a patient with aMCI is described as being in a 'prodromal stage of Alzheimer's disease' or as having 'MCI due to Alzheimer's disease', she and her relatives may immediately fear and probably also plan for a future with severe symptoms of dementia. Considering the uncertainty related to who will show a progressive decline on an individual basis, as well as the current lack of effective treatments, it is of utmost importance to carefully consider how to communicate information about risk, as well as resilience factors and coping strategies. As shown in the present study, aMCI at an early stage is not always synonymous with a trajectory towards AD.

## 4.1 Strengths and Limitations

There are several strengths of the present study. The results were made possible by including a large cohort of elderly with aMCI from the open ADNI dataset, by using longitudinal diagnostic labels to identify the two MCI subgroups, and by an analytic approach within a machine learning framework. This strength was dependent on a substantial effort put into data preparation before conducting the statistical analyses. Despite one main objective of the ADNI project to make their data openly available, the structure of the ADNI database makes it challenging to select subsets of longitudinal data. This is related to, but not restricted to, their inclusion of the four different study phases, each with somewhat different schedules of visits and study protocols. As a consequence, data for subjects in one phase is stored under one file name, while equivalent data for subjects from other phases are stored in another file, often with different naming conventions for the same variables. A substantial proportion of studies employing ADNI data have therefore either i) restricted inclusion of subjects to one or two study phases, ii) used ADNI's pre-prepared data set (`ADNIMERGE.csv`), which is more easily accessible at the cost of having a restricted selection of variables, or iii) dealt with a high number of missing values. By the effort put into data preparation in the present study, we enabled inclusion of a relatively large sample size with few missing values on a set of clinically relevant cognitive, genetic, and MRI data. As such, we obtained a sample size large enough to set aside a test set for final evaluation, while still preserving a sufficiently high number of participants (i.e. enough data) for the model to be trained on. Furthermore, selecting participants with a long follow-up time decreased the chance of wrongfully labeling individuals who would have progressed to AD, given more time passing, as stable MCI. Finally, the effort to re-analyze MRI data with the most recent version of FreeSurfer, rather than using the older FreeSurfer outputs made available as part of the ADNI dataset, should also be mentioned as an important strength.

Some limitations must nevertheless be stated. First and foremost our sample consisted of subjects who in ADNI received a baseline diagnosis of MCI. As the cognitive decline related to MCI typically has an insidious onset with no fixed events defining its start, it is often challenging to temporally anchor the point of transition. The baseline assessment was used as a proxy for the initial examination a person in the 'real world' would receive if reporting concerns about cognitive dysfunction. Nevertheless, having MCI at the baseline visit of ADNI does not necessarily mean that subjects were cognitively non-impaired prior to enrollment. One way of being sure to exclusively capture the initial phases of the MCI stage could be to include subjects who had normal cognition at the time of enrollment, and subsequently converted to MCI. Then, the visit in which the subject was first characterized as MCI could be used as baseline. This would however result in drastic reduction of the sample size. The homogeneous sample included in the ADNI dataset is another limitation. The sample mainly includes highly educated and motivated volunteers geographically restricted to North-America. Given this, additional studies relying on other samples are required to assess how the findings of the current study generalises to other populations.

## 4.2 Future Research

MCI is a multidetermined diagnostic entity, where several interacting environmental and biological factors contribute to individual differences in clinical outcome. This variability suggests the existence of underlying subphenotypes driven by different pathophysiological mechanisms leading to a similar clinical outcome, referred to as equifinality by Fezcko and colleagues (2019). Several studies have tried to untangle the heterogeneity by the use of hypothesis-driven designs (see for instance Byun et al., 2015; Risacher et al., 2017 for studies employing ADNI data). Such approaches can reveal important information on specific characteristics of these subtypes, but they are nevertheless restricted to a priori

definitions. More recently, empirically driven studies have illustrated that there may be great heterogeneity even within the aMCI subgroup, both with respect to cognitive profiles (Edmonds et al., 2014) and patterns of cortical atrophy (Edmonds et al., 2016). To more directly address this heterogeneity, future studies should build on the current framework by incorporating the *functional* Random Forest model proposed by Fezcko and colleagues (2020). Through this hybrid approach, the supervised RF is followed by an unsupervised community detection algorithm used to identify putative subgroups in the MCI population. The subsequent employment of this unsupervised method means that no assumptions regarding the number or nature of subgroups have to be made a priori. This is important in the context of precision medicine, as it could potentially yield valuable new insights about subgroups of individuals on a trajectory towards AD. Plans for such a study are already established, and the author of this thesis has presented these plans in a poster presented at a National conference (Rye et al., 2020). In future studies, a wider range of features should also be included to cover both risk and resilience factors. Furthermore, information about longitudinal change should be taken into account in the analytic model, as the rate of decline is expected to be more dramatic as the patient is closer to a definite neurodegenerative disease.

# 5 Conclusion

AD is a fatal disorder with a huge impact on the lives of those affected and their caregivers. In addition to personal consequences, the economical costs related to the disease are massive. With increased longevity, and age being the primary risk factor for developing AD, the world is facing what has been described as an epidemic related to this neurodegenerative disease in the coming years. The need for new and effective treatments therefore calls for immediate action. Nevertheless, the continuous lack of successful drug trials indicates that novel approaches are needed, and a pressing issue related to this is the early and reliable identification of MCI subjects who are on a trajectory towards AD.

This thesis has contributed to this in several ways. Firstly, by illustrating that differences in both biological markers and clinical phenotype between aMCI subjects who remain stable and those converting to AD could be identified already at a baseline assessment. Secondly, by constructing a Random Forest machine learning algorithm trained on baseline data, we predicted conversion with an accuracy much better than chance level. Lastly, we extended this latter contribution by 'looking inside the machine learning black box' to identify what specific features were most important for making this prediction, as well as how specific values of these features affected risk of conversion.

Importantly, this thesis has also shed light on the heterogeneity and complexity characterizing the MCI construct, including a variety of etiologically and neuropathologically distinct conditions resulting in differing patterns of cognitive impairments and developmental trajectories. For early identification and effective treatment of AD, each individual risk profile should be taken into account. To achieve this, large amounts of high-dimensional data from several modalities needs to be considered. As it is extremely challenging for the human mind (i.e. health professionals) to identify patterns in such high dimensional data, we strongly believe that interdisciplinary cooperation combining clinical

and computational expertise is extremely valuable. The current study is one small contribution to this, showing that we seem to be right in the middle of a paradigm shift in the way we conceptualize, diagnose and treat AD; moving away from a "one-size-fits-all" towards a precision medicine approach.

## References

Adólfsdóttir, S., Wollschlaeger, D., Wehling, E., & Lundervold, A. J. (2017). Inhibition and switching in healthy aging: A longitudinal study. *Journal of the International Neuropsychological Society*, *23*(1), 90–97. https://doi.org/10.1017/s1355617716000898

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 270–279. https://doi.org/10.1016/j.jalz.2011.03.008

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders (dsm-5)*. (5th ed.) American Psychiatric Association Publishing. https://www.ebook.de/de/product/20229522/american_psychiatric_association_diagnostic_and_statistical_manual_of_mental_disorders_dsm_5_r.html

Amoroso, N., Diacono, D., Fanizzi, A., Rocca, M. L., Monaco, A., Lombardi, A., Guaragnella, C., Bellotti, R., & Tangaro, S. (2018). Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge. *Journal of Neuroscience Methods*, *302*, 3–9. https://doi.org/10.1016/j.jneumeth.2017.12.011

Apostolova, L. G., Green, A. E., Babakchanian, S., Hwang, K. S., Chou, Y.-Y., Toga, A. W., & Thompson, P. M. (2012). Hippocampal Atrophy and Ventricular Enlargement in Normal Aging, Mild Cognitive Impairment (MCI), and Alzheimer Disease. *Alzheimer Disease & Associated Disorders*, *26*(1), 17–27. https://doi.org/10.1097/wad.0b013e3182163b62

Arenaza-Urquijo, E. M., & Vemuri, P. (2018). Resistance vs resilience to Alzheimer disease. *Neurology*, *90*(15), 695–703. https://doi.org/10.1212/wnl.0000000000005303

Baldo, J. V., & Shimamura, A. P. (1998). Letter and Category Fluency in Patients with Frontal Lobe Lesions. *Neuropsychology*, *12*(2), 259–267. https://doi.org/10.1037/0894-4105.12.2.259

Battista, P., Salvatore, C., & Castiglioni, I. (2017). Optimizing Neuropsychological Assessments for Cognitive, Behavioral, and Functional Impairment Classification: A Machine Learning Study. *Behavioural Neurology*, *2017*, 1–19. https://doi.org/10.1155/2017/1850909

Berkowitz, C., Mosconi, L., Scheyer, O., Rahman, A., Hristov, H., & Isaacson, R. (2018). Precision Medicine for Alzheimer's Disease Prevention. *Healthcare*, *6*(3), 1–11. https://doi.org/10.3390/healthcare6030082

Berkowitz, C., Mosconi, L., Rahman, A., Scheyer, O., Hristov, H., & Isaacson, R. (2018). Clinical Application of APOE in Alzheimer's Prevention: A Precision Medicine Approach. *The Journal Of Prevention of Alzheimer's Disease*, 245–252. https://doi.org/10.14283/jpad.2018.35

Blazer, D. (2013). Neurocognitive Disorders in DSM-5. *American Journal of Psychiatry*, *170*(6), 585–587. https://doi.org/10.1176/appi.ajp.2013.13020179

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the trail making test. *Nature Protocols*, *1*(5), 2277–2281. https://doi.org/10.1038/nprot.2006.390

Braak, H., & Braak, E. (1991). Neuropathological stageing of alzheimer-related changes. *Acta Neuropathologica*, *82*(4), 239–259. https://doi.org/https://doi.org/10.1007/BF00308809

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brevik, E. J., Eikeland, R. A., & Lundervold, A. J. (2013). Subthreshold depressive symptoms have a negative impact on cognitive functioning in middle-aged and older males. *Frontiers in Psychology*, *4*, 309. https://doi.org/10.3389/fpsyg.2013.00309

Brown, E. M., Pierce, M. E., Clark, D. C., Fischl, B. R., Iglesias, J. E., Milberg, W. P., McGlinchey, R. E., & Salat, D. H. (2020). Test-retest reliability of FreeSurfer automated hippocampal subfield segmentation within and across scanners. *NeuroImage*, *210*, 1–12. https://doi.org/10.1016/j.neuroimage.2020.116563

Butt, O. H., Meeker, K. L., Wisch, J. K., Schindler, S. E., Fagan, A. M., Benzinger, T. L., Cruchaga, C., Holtzman, D. M., Morris, J. C., & Ances, B. M. (2021). Network dysfunction in cognitively normal APOE$\varepsilon$4 carriers is related to subclinical tau. *Alzheimer's & Dementia*, 1–11. https://doi.org/10.1002/alz.12375

Butters, N., Granholm, E., Salmon, D. P., Grant, I., & Wolfe, J. (1987). Episodic and Semantic Memory: A Comparison of Amnesic and Demented Patients. *Journal of Clinical and Experimental Neuropsychology*, *9*(5), 479–497. https://doi.org/10.1080/01688638708410764

Byun, M. S., Kim, S. E., Park, J., Yi, D., Choe, Y. M., Sohn, B. K., Choi, H. J., Baek, H., Han, J. Y., Woo, J. I., & and, D. Y. L. (2015). Heterogeneity of Regional Brain Atrophy Patterns Associated with Distinct Progression Rates in Alzheimer's Disease (D.-G. Jo, Ed.). *PLOS ONE*, *10*(11), e0142756. https://doi.org/10.1371/journal.pone.0142756

Caillaud, M., Hudon, C., Boller, B., Brambati, S., Duchesne, S., Lorrain, D., Gagnon, J.-F., Maltezos, S., Mellah, S., Phillips, N., & and, S. B. (2019). Evidence of a Relation Between Hippocampal Volume, White Matter Hyperintensities, and Cognition in

Subjective Cognitive Decline and Mild Cognitive Impairment (A. Gutchess, Ed.). *The Journals of Gerontology: Series B*, *75*(7), 1382–1392. https://doi.org/10.1093/geronb/gbz120

Calabro, M., Rinaldi, C., Santoro, G., & Crisafulli, C. (n.d.). The biological pathways of Alzheimer disease: a review, year = 2021. *AIMS Neuroscience*, *8*(1), 86–132. https://doi.org/10.3934/neuroscience.2021005

Carreiro, A. V., Mendonça, A., de Carvalho, M., & Madeira, S. C. (2015). Integrative biomarker discovery in neurodegenerative diseases. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *7*(6), 357–379. https://doi.org/10.1002/wsbm.1310

Castellano, J. M., Kim, J., Stewart, F. R., Jiang, H., DeMattos, R. B., Patterson, B. W., Fagan, A. M., Morris, J. C., Mawuenyega, K. G., Cruchaga, C., Goate, A. M., Bales, K. R., Paul, S. M., Bateman, R. J., & Holtzman, D. M. (2011). Human apoE Isoforms Differentially Regulate Brain Amyloid-$\beta$ Peptide Clearance. *Science Translational Medicine*, *3*(89), 89ra57. https://doi.org/10.1126/scitranslmed.3002156

Cauwenberghe, C. V., Broeckhoven, C. V., & Sleegers, K. (2015). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in Medicine*, *18*(5), 421–430. https://doi.org/10.1038/gim.2015.117

Chasioti, D., Yan, J., Nho, K., & Saykin, A. J. (2019). Progress in Polygenic Composite Scores in Alzheimer's and Other Complex Diseases. *Trends in Genetics*, *35*(5), 371–382. https://doi.org/10.1016/j.tig.2019.02.005

Crystal, H., Dickson, D., Fuld, P., Masur, D., Scott, R., Mehler, M., Masdeu, J., Kawas, C., Aronson, M., & Wolfson, L. (1988). Clinico-pathologic studies in dementia: Nondemented subjects with pathologically confirmed alzheimer's disease. *Neurology*, *38*(11), 1682–1687. https://doi.org/10.1212/wnl.38.11.1682

Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J., & Anderberg, P. (2017). Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review (K. Chen, Ed.). *PLOS ONE*, *12*(6), e0179804. https://doi.org/10.1371/journal.pone.0179804

Desikan, R. S., Fan, C. C., Wang, Y., Schork, A. J., Cabral, H. J., Cupples, L. A., Thompson, W. K., Besser, L., Kukull, W. A., Holland, D., Chen, C.-H., Brewer, J. B., Karow, D. S., Kauppi, K., Witoelar, A., Karch, C. M., Bonham, L. W., Yokoyama, J. S., Rosen, H. J., … Dale, A. M. (2017). Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score (C. Brayne, Ed.). *PLOS Medicine*, *14*(3), e1002258. https://doi.org/10.1371/journal.pmed.1002258

Dubois, B., Picard, G., & Sarazin, M. (2009). Early detection of Alzheimer's disease: new diagnostic criteria. *Alzheimer's Disease and Mild Cognitive Impairment*, *11*(2), 135–139. https://doi.org/10.31887/dcns.2009.11.2/bdubois

Edmonds, E. C., Delano-Wood, L., Clark, L. R., Jak, A. J., Nation, D. A., McDonald, C. R., Libon, D. J., Au, R., Galasko, D., Salmon, D. P., & and, M. W. B. (2014). Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors. *Alzheimer's & Dementia*, *11*(4), 415–424. https://doi.org/10.1016/j.jalz.2014.03.005

Edmonds, E. C., Eppig, J., Bondi, M. W., Leyden, K. M., Goodwin, B., Delano-Wood, L., & and, C. R. M. (2016). Heterogeneous cortical atrophy patterns in MCI not captured by conventional diagnostic criteria. *Neurology*, *87*(20), 2108–2116. https://doi.org/10.1212/wnl.0000000000003326

Feczko, E., & Fair, D. A. (2020). Methods and Challenges for Assessing Heterogeneity. *Biological Psychiatry*, *88*(1), 9–17. https://doi.org/10.1016/j.biopsych.2020.02.015

Feczko, E., Miranda-Dominguez, O., Marr, M., Graham, A. M., Nigg, J. T., & Fair, D. A. (2019). The Heterogeneity Problem: Approaches to Identify Psychiatric Subtypes. *Trends in Cognitive Sciences*, *23*(7), 584–601. https://doi.org/10.1016/j.tics.2019.03.009

Frankó, E., & and, O. J. (2013). Evaluating Alzheimer's Disease Progression Using Rate of Regional Hippocampal Atrophy (K. Herholz, Ed.). *PLoS ONE*, *8*(8), e71354. https://doi.org/10.1371/journal.pone.0071354

Fulton-Howard, B., Goate, A. M., Adelson, R. P., Koppel, J., Gordon, M. L., Barzilai, N., Atzmon, G., Davies, P., & Freudenberg-Hua, Y. (2021). Greater Effect of Polygenic Risk Score for Alzheimer's Disease Among Younger Cases who are Apolipoprotein E-$\varepsilon$4 Carriers. *Neurobiology of Aging*, *99*, 1–9. https://doi.org/https://doi.org/10.1016/j.neurobiolaging.2020.09.014

Galton, C. J., Patterson, K., Graham, K., Lambon-Ralph, M., Williams, G., Antoun, N., Sahakian, B., & Hodges, J. (2001). Differing patterns of temporal atrophy in Alzheimer's disease and semantic dementia. *Neurology*, *57*(2), 216–225. https://doi.org/10.1212/wnl.57.2.216

Gauthier, S., Reisberg, B., Zaudig, M., Petersen, R. C., Ritchie, K., Broich, K., Belleville, S., Brodaty, H., Bennett, D., Chertkow, H., Cummings, J. L., de Leon, M., Feldman, H., Ganguli, M., Hampel, H., Scheltens, P., Tierney, M. C., Whitehouse, P., & Winblad, B. (2006). Mild cognitive impairment. *The Lancet*, *367*(9518), 1262–1270. https://doi.org/10.1016/s0140-6736(06)68542-5

Gorbach, T., Pudas, S., Bartrés-Faz, D., Brandmaier, A. M., Düzel, S., Henson, R. N., Idland, A.-V., Lindenberger, U., Bros, D. M., Mowinckel, A. M., Solé-Padullés, C., Sørensen, Ø., Walhovd, K. B., Watne, L. O., Westerhausen, R., Fjell, A. M., & Nyberg, L. (2020). Longitudinal association between hippocampus atrophy and episodic-memory decline in non-demented APOE $\varepsilon$4 carriers. *Alzheimer's & De-*

*mentia: Diagnosis, Assessment & Disease Monitoring*, *12*(1), 1–9. https://doi.org/10.1002/dad2.12110

Guarino, A., Favieri, F., Boncompagni, I., Agostini, F., Cantone, M., & Casagrande, M. (2019). Executive Functions in Alzheimer Disease: A Systematic Review. *Frontiers in Aging Neuroscience*, *10*, 437–461. https://doi.org/10.3389/fnagi.2018.00437

Gutchess, A. (2019). *Cognitive and social neuroscience of aging*. Cambridge.

Hampel, H., Prvulovic, D., Teipel, S., Jessen, F., Luckhaus, C., Frölich, L., Riepe, M. W., Dodel, R., Leyhe, T., Bertram, L., Hoffmann, W., & Faltraco, F. (2011). The future of Alzheimer's disease: The next 10 years. *Progress in Neurobiology*, *95*(4), 718–728. https://doi.org/10.1016/j.pneurobio.2011.11.008

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. (2nd ed.). Springer.

Holtzman, D. M., Morris, J. C., & Goate, A. M. (2011). Alzheimer's Disease: The Challenge of the Second Century. *Science Translational Medicine*, *3*(77), 1–17. https://doi.org/10.1126/scitranslmed.3002369

Huang, L.-K., Chao, S.-P., & Hu, C.-J. (2020). Clinical trials of new drugs for Alzheimer disease. *Journal of Biomedical Science*, *27*(1), 1–13. https://doi.org/10.1186/s12929-019-0609-7

Iqbal, K., Alonso, A. D., Chen, S., Chohan, M. O., El-Akkad, E., Gong, C.-X., Khatoon, S., Li, B., Liu, F., Rahman, A., Tanimukai, H., & Grundke-Iqbal, I. (2005). Tau pathology in Alzheimer disease and other tauopathies. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, *1739*(2-3), 198–210. https://doi.org/10.1016/j.bbadis.2004.09.008

Jack, C. R., Wiste, H. J., Vemuri, P., Weigand, S. D., Senjem, M. L., Zeng, G., Bernstein, M. A., Gunter, J. L., Pankratz, V. S., Aisen, P. S., Weiner, M. W., Petersen, R. C., Shaw, L. M., Trojanowski, J. Q., & and, D. S. K. (2010). Brain beta-amyloid mea-

sures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimer's disease. *Brain*, *133*(11), 3336–3348. https://doi.org/10.1093/brain/awq277

Jack, C. R., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., Thies, B., & Phelps, C. H. (2011). Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 257–262. https://doi.org/10.1016/j.jalz.2011.03.004

Jack, C. R., & Holtzman, D. M. (2013). Biomarker Modeling of Alzheimer's Disease. *Neuron*, *80*(6), 1347–1358. https://doi.org/10.1016/j.neuron.2013.12.003

Japkowicz, N., & Shah, M. (2009). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press.

Khan, W., Westman, E., Jones, N., Wahlund, L.-O., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Spenger, C., Lovestone, S., Muehlboeck, J.-S., & Simmons, A. (2014). Automated Hippocampal Subfield Measures as Predictors of Conversion from Mild Cognitive Impairment to Alzheimer's Disease in Two Independent Cohorts. *Brain Topography*, *28*(5), 746–759. https://doi.org/10.1007/s10548-014-0415-1

Koepsell, T. D., & Monsell, S. E. (2012). Reversion from mild cognitive impairment to normal or near-normal cognition: Risk factors and prognosis. *Neurology*, *79*(15), 1591–1598. https://doi.org/10.1212/wnl.0b013e31826e26b7

Kryscio, R. J., Schmitt, F. A., Salazar, J. C., Mendiondo, M. S., & Markesbery, W. R. (2006). Risk factors for transitions from normal to mild cognitive impairment and dementia. *Neurology*, *66*(6), 828–832. https://doi.org/10.1212/01.wnl.0000203264.71880.45

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment*. Oxford University Press.

Liu, C.-C., Kanekiyo, T., Xu, H., & Bu, G. (2013). Apolipoprotein e and alzheimer disease: Risk, mechanisms and therapy. *Nature Reviews Neurology*, *9*(2), 106–118. https://doi.org/10.1038/nrneurol.2012.263

Lundervold, A. J., Vik, A., & Lundervold, A. (2019). Lateral ventricle volume trajectories predict response inhibition in older age—a longitudinal brain imaging and machine learning approach. *PLOS ONE*, *14*(4), e0207967. https://doi.org/10.1371/journal.pone.0207967

Lyketsos, C. G., Lopez, O., Jones, B., Fitzpatrick, A. L., Breitner, J., & DeKosky, S. (2002). Prevalence of Neuropsychiatric Symptoms in Dementia and Mild Cognitive Impairment. *JAMA*, *288*(12), 1475–1483. https://doi.org/10.1001/jama.288.12.1475

Martin, E., & Velayudhan, L. (2020). Neuropsychiatric Symptoms in Mild Cognitive Impairment: A Literature Review. *Dementia and Geriatric Cognitive Disorders*, *49*(2), 146–155. https://doi.org/10.1159/000507078

McCabe, D. P., III, H. L. R., McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember–know judgments. *Neuropsychologia*, *47*(11), 2164–2173. https://doi.org/10.1016/j.neuropsychologia.2008.11.025

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's diseaseReport of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*(7), 939–944. https://doi.org/10.1212/wnl.34.7.939

Modrego, P. J., & Ferrández, J. (2004). Depression in Patients With Mild Cognitive Impairment Increases the Risk of Developing Dementia of Alzheimer Type. *Archives of Neurology*, *61*(8), 1290–1293. https://doi.org/10.1001/archneur.61.8.1290

Mofrad, S. A., Lundervold, A., & Lundervold, A. S. (2021). A predictive framework based on brain volume trajectories enabling early detection of Alzheimer's disease. *Computerized Medical Imaging and Graphics*, *90*, 101910. https://doi.org/10.1016/j.compmedimag.2021.101910

Molano, J., Boeve, B., Ferman, T., Smith, G., Parisi, J., Dickson, D., Knopman, D., Graff-Radford, N., Geda, Y., Lucas, J., Kantarci, K., Shiung, M., Jack, C., Silber, M., Pankratz, V. S., & Petersen, R. (2009). Mild cognitive impairment associated with limbic and neocortical lewy body disease: A clinicopathological study. *Brain*, *133*(2), 540–556. https://doi.org/10.1093/brain/awp280

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *NeuroImage*, *104*, 398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002

Morris, J. C., Storandt, M., McKeel, D. W., Rubin, E. H., Price, J. L., Grant, E. A., & Berg, L. (1996). Cerebral amyloid deposition and diffuse plaques in "normal" aging: Evidence for presymptomatic and very mild Alzheimer's disease. *Neurology*, *46*(3), 707–719. https://doi.org/10.1212/wnl.46.3.707

Naqa, I. E., Li, R., & Murphy, M. J. (2015). *Machine learning in radiation oncology: Theory and applications*. Springer International Publishing.

Nelson, H. E., & O'Connell, A. (1978). Dementia: The Estimation of Premorbid Intelligence Levels Using the New Adult Reading Test. *Cortex*, *14*(2), 234–244. https://doi.org/10.1016/s0010-9452(78)80049-5

Neuropathology Group. (2001). Pathological correlates of late-onset dementia in a multicentre, community-based population in England and Wales. *The Lancet*, *357*(9251), 169–175. https://doi.org/10.1016/s0140-6736(00)03589-3

Nobis, L., Manohar, S. G., Smith, S. M., Alfaro-Almagro, F., Jenkinson, M., Mackay, C. E., & Husain, M. (2019). Hippocampal volume across age: Nomograms derived from over 19,700 people in UK biobank. *NeuroImage: Clinical*, *23*, 1–13. https://doi.org/10.1016/j.nicl.2019.101904

Nyberg, L., Lövdén, M., Riklund, K., Lindenberger, U., & Bäckman, L. (2012). Memory aging and brain maintenance. *Trends in Cognitive Sciences*, *16*(5), 292–305. https://doi.org/10.1016/j.tics.2012.04.005

Nyberg, L., & Pudas, S. (2019). Successful Memory Aging. *Annual Review of Psychology*, *70*(1), 219–243. https://doi.org/10.1146/annurev-psych-010418-103052

Oh, H., Madison, C., Haight, T. J., Markley, C., & Jagust, W. J. (2012). Effects of age and beta-amyloid on cognitive changes in normal elderly people. *Neurobiology of Aging*, *33*(12), 2746–2755. https://doi.org/10.1016/j.neurobiolaging.2012.02.008

Palmer, K., Iulio, F. D., Varsi, A. E., Gianni, W., Sancesario, G., Caltagirone, C., & Spalletta, G. (2010). Neuropsychiatric Predictors of Progression from Amnestic-Mild Cognitive Impairment to Alzheimer's Disease: The Role of Depression and Apathy. *Journal of Alzheimer's Disease*, *20*, 175–183. https://doi.org/10.3233/JAD-2010-1352

Panza, F., Lozupone, M., Solfrizzi, V., Sardone, R., Dibello, V., Lena, L. D., D'Urso, F., Stallone, R., Petruzzi, M., Giannelli, G., Quaranta, N., Bellomo, A., Greco, A., Daniele, A., Seripa, D., & Logroscino, G. (2018). Different Cognitive Frailty Models and Health- and Cognitive-related Outcomes in Older Age: From Epidemiology to Prevention. *Journal of Alzheimer's Disease*, *62*(3), 993–1012. https://doi.org/10.3233/JAD-170963

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of Visuospatial and Verbal Memory Across the Adult Life Span. *Psychology and Aging*, *17*(2), 299–320. https://doi.org/10.1037/0882-7974.17.2.299

Pereira, T., Ferreira, F. L., Cardoso, S., Silva, D., de Mendonca, A., Guerreiro, M., & Madeira, S. C. (2018). Neuropsychological predictors of conversion from mild cognitive impairment to Alzheimer's disease: a feature selection ensemble combining stability and predictability. *BMC Medical Informatics and Decision Making*, *18*(1), 1–20. https://doi.org/10.1186/s12911-018-0710-y

Petersen, R. C. (2004a). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, *256*(3), 183–194. https://doi.org/10.1111/j.1365-2796.2004.01388.x

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., Harvey, D. J., Jack, C. R., Jagust, W. J., Shaw, L. M., Toga, A. W., Trojanowski, J. Q., & Weiner, M. W. (2009). Alzheimer's disease neuroimaging initiative (ADNI): Clinical characterization. *Neurology*, *74*(3), 201–209. https://doi.org/10.1212/wnl.0b013e3181cb3e25

Petersen, R. C., Jack, C. R., Xu, Y.-C., Waring, S. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J., Tangalos, E. G., Boeve, B. F., & Kokmen, E. (2000). Memory and MRI-based hippocampal volumes in aging and AD. *Neurology*, *54*(3), 581–587. https://doi.org/10.1212/wnl.54.3.581

Petersen, R. C. (2004b). Mild cognitive impairment. *CONTINUUM: Lifelong Learning in Neurology*, *10*, 9–28. https://doi.org/10.1212/01.con.0000293545.39683.cc

Petersen, R. C. (2011). Mild cognitive impairment. *New England Journal of Medicine*, *364*(23), 2227–2234. https://doi.org/10.1056/nejmcp0910237

Petersen, R. C., Parisi, J. E., Dickson, D. W., Johnson, K. A., Knopman, D. S., Boeve, B. F., Jicha, G. A., Ivnik, R. J., Smith, G. E., Tangalos, E. G., Braak, H., & Kok-

men, E. (2006). Neuropathologic Features of Amnestic Mild Cognitive Impairment. *Archives of Neurology*, *63*(5), 665–672. https://doi.org/10.1001/archneur.63.5.665

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G., & Kokmen, E. (1999). Mild cognitive impairment: Clinical characterization and outcome. *Archives of Neurology*, *56*(3), 303–308. https://doi.org/10.1001/archneur.56.3.303

Prince, M. J., Wu, F., Guo, Y., Robledo, L. M. G., O'Donnell, M., Sullivan, R., & Yusuf, S. (2015). The burden of disease in older people and implications for health policy and practice. *The Lancet*, *385*(9967), 549–562. https://doi.org/10.1016/s0140-6736(14)61347-7

Purves, D., Cabeza, R., Huettel, S. A., LaBar, K. S., Platt, M. L., & Waldorff, M. G. (2013). *Principles of cognitive neuroscience*. (2nd ed.). Sinauer Associates.

Rahman, M. A., Rahman, M. S., Uddin, M. J., Mamum-Or-Rashid, A. N. M., Pang, M.-G., & Rhim, H. (2020). Emerging risk of environmental factors: insight mechanisms of Alzheimer's diseases. *Environmental Science and Pollution Research*, *27*(36), 44659–44672. https://doi.org/10.1007/s11356-020-08243-z

Raz, N., & Kennedy, K. M. (2009). A systems approach to the aging brain: Neuroanatomic changes, their modifiers, and cognitive correlates. *Imaging the aging brain* (pp. 43–70). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195328875.003.0004

Reitan, R. M. (1958). Validity of the trail making test indicator of organic brain damage. *Percept Mot Skills*, *8*(7), 271–276. https://doi.org/https://doi.org/10.2466/pms.1958.8.3.271

Reitz, C., Rogaeva, E., & Beecham, G. W. (2020). Late-onset vs nonmendelian early-onset Alzheimer disease. *Neurology Genetics*, *6*(5), 1–9. https://doi.org/10.1212/nxg.0000000000000512

Reuter, M., & Fischl, B. (2011). Avoiding asymmetry-induced bias in longitudinal image processing. *NeuroImage*, *57*(1), 19–21. https://doi.org/10.1016/j.neuroimage.2011.02.076

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: A robust approach. *NeuroImage*, *53*(4), 1181–1196. https://doi.org/10.1016/j.neuroimage.2010.07.020

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, *61*(4), 1402–1418. https://doi.org/10.1016/j.neuroimage.2012.02.084

Rey, A. (1964). *Cognitive and social neuroscience of aging*. Universitaires de France.

Ricciarelli, R., & Fedele, E. (2017). The Amyloid Cascade Hypothesis in Alzheimer's Disease: It's Time to Change Our Mind. *Current Neuropharmacology*, *15*(6), 926–935. https://doi.org/10.2174/1570159x15666170116143743

Risacher, S. L., Anderson, W. H., Charil, A., Castelluccio, P. F., Shcherbinin, S., Saykin, A. J., Schwarz, A. J., & Alzheimer's Disease Neuroimaging Initiative. (2017). Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline. *Neurology*, *89*(21), 2176–2186. https://doi.org/10.1212/wnl.0000000000004670

Roda, A. R., Montoliu-Gaya, L., & Villegas, S. (2019). The Role of Apolipoprotein E Isoforms in Alzheimer's Disease. *Journal of Alzheimer's Disease*, *68*(2), 459–471. https://doi.org/10.3233/JAD-180740

Rogalski, E. J., Gefen, T., Shi, J., Samimi, M., Bigio, E., Weintraub, S., Geula, C., & Mesulam, M.-M. (2013). Youthful Memory Capacity in Old Brains: Anatomic and Genetic Clues from the Northwestern SuperAging Project. *Journal of Cognitive Neuroscience*, *25*(1), 29–36. https://doi.org/10.1162/jocn_a_00300

Rozzini, L., Chilovi, B. V., Conti, M., Delrio, I., Borroni, B., Trabucchi, M., & Padovani, A. (2007). Neuropsychiatric Symptoms in Amnestic and Nonamnestic Mild Cognitive Impairment. *Dementia and Geriatric Cognitive Disorders*, *25*(1), 32–36. https://doi.org/10.1159/000111133

Rye, I., Kocinski, M., Vik, A., Lundervold, A. J., & Lundervold, A. S. (2020). Mci-subgrups. https://github.com/ingryy/mci_subgrups/blob/main/poster_conference2020.pdf

Sachs-Ericsson, N., & Blazer, D. G. (2014). The new DSM-5 diagnosis of mild neurocognitive disorder and its relation to research in mild cognitive impairment. *Aging & Mental Health*, *19*(1), 2–12. https://doi.org/10.1080/13607863.2014.920303

Samaranch, L., Cervantes, S., Barabash, A., Alonso, A., Cabranes, J. A., Lamet, I., Ancín, I., Lorenzo, E., Martínez-Lage, P., Marcos, A., Clarimón, J., Alcolea, D., Lleó, A., Blesa, R., Gómez-Isla, T., & Pastor, P. (2011). The Effect of MAPT H1 and APOE 4 on Transition from Mild Cognitive Impairment to Dementia. *Journal of Alzheimer's Disease*, *22*(4), 1065–1071. https://doi.org/10.3233/JAD-2010-101011

Schwartz, S., & Baldo, J. (2001). Distinct patterns of word retrieval in right and left frontal lobe patients: A multidimensional perspective. *Neuropsychologia*, *39*(11), 1209–1217. https://doi.org/10.1016/s0028-3932(01)00053-7

Serrano-Pozo, A., Frosch, M. P., Masliah, E., & Hyman, B. T. (2011). Neuropathological Alterations in Alzheimer Disease. *Cold Spring Harbor Perspectives in Medicine*, *1*(1), 1–23. https://doi.org/10.1101/cshperspect.a006189

Shao, Z., Janse, E., Visser, K., & Meyer, A. S. (2014). What do verbal fluency tasks measure? Predictors of verbal fluency performance in older adults. *Frontiers in Psychology*, *5*, 1–10. https://doi.org/10.3389/fpsyg.2014.00772

Sims, R., Hill, M., & Williams, J. (2020). The multiplex model of the genetics of Alzheimer's disease. *Nature Neuroscience*, *23*(3), 311–322. https://doi.org/10.1038/s41593-020-0599-5

Skogli, E., Karttinen, E., Stokke, O. M., & Vikøren, S. (2020). *Samfunskostnader knyttet til Alzheimers og annen demenssykdom* (tech. rep.). MENON-PUBLIKASJON NR. 64/2020. https://doi.org/https://www.menon.no/publication/samfunnskostnader-knyttet-alzheimers-annen-demens/

Sperling, R. A., Jack, C. R., & Aisen, P. S. (2011). Testing the Right Target and Right Drug at the Right Stage. *Science Translational Medicine*, *3*(111), 1–5. https://doi.org/10.1126/scitranslmed.3002609

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack, C. R., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., … Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia*, *7*(3), 280–292. https://doi.org/10.1016/j.jalz.2011.03.003

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 1–11. https://doi.org/10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 1–21. https://doi.org/10.1186/1471-2105-8-25

ten Kate, M., Dicks, E., Visser, P. J., van der Flier, W. M., Teunissen, C. E., Barkhof, F., Scheltens, P., & and, B. M. T. (2018). Atrophy subtypes in prodromal Alzheimer's

disease are associated with cognitive decline. *Brain*, *141*(12), 3443–3456. https://doi.org/10.1093/brain/awy264

Tervo, S., Kivipelto, M., Hänninen, T., Vanhanen, M., Hallikainen, M., Mannermaa, A., & Soininen, H. (2004). Incidence and Risk Factors for Mild Cognitive Impairment: A Population-Based Three-Year Follow-Up Study of Cognitively Healthy Elderly Subjects. *Dementia and Geriatric Cognitive Disorders*, *17*(3), 196–203. https://doi.org/10.1159/000076356

Tucker, A. M., & Stern, Y. (2011). Cognitive Reserve in Aging. *Current Alzheimer Research*, *8*(4), 354–360. https://doi.org/10.2174/156720511795745320

Varatharajah, Y., Ramanan, V. K., Iyer, R., & Vemuri, P. (2019). Predicting Short-term MCI-to-AD Progression Using Imaging, CSF, Genetic Factors, Cognitive Resilience, and Demographics. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-38793-3

Vasta, R., Augimeri, A., Cerasa, A., Nigro, S., Gramigna, V., Nonnis, M., Rocca, F., Zito, G., Quattrone, A., & the Alzheimer's Disease Neuroimaging. (2016). Hippocampal Subfield Atrophies in Converted and Not-Converted Mild Cognitive Impairments Patients by a Markov Random Fields Algorithm. *Current Alzheimer Research*, *13*(5), 566–574. https://doi.org/10.2174/1567205013666160120151457

Walhovd, K. B., Fjell, A. M., & Espeseth, T. (2014). Cognitive decline and brain pathology in aging - need for a dimensional, lifespan and systems vulnerability view. *Scandinavian Journal of Psychology*, *55*(3), 244–254. https://doi.org/10.1111/sjop.12120

Winblad, B., Palmer, K., Kivipelto, M., Jelic, V., Fratiglioni, L., Wahlund, L.-O., Nordberg, A., Backman, L., Albert, M., Almkvist, O., Arai, H., Basun, H., Blennow, K., de Leon, M., DeCarli, C., Erkinjuntti, T., Giacobini, E., Graff, C., Hardy, J., ... Petersen, R. (2004). Mild cognitive impairment - beyond controversies, towards a consensus: report of the International Working Group on Mild Cognitive Impairment.

*Journal of Internal Medicine*, *256*(3), 240–246. https://doi.org/10.1111/j.1365-2796.2004.01380.x

Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., Cedazo-Minguez, A., Dubois, B., Edvardsson, D., Feldman, H., Fratiglioni, L., Frisoni, G. B., Gauthier, S., Georges, J., Graff, C., Iqbal, K., Jessen, F., Johansson, G., Jönsson, L., … Zetterberg, H. (2016). Defeating Alzheimer's disease and other dementias: a priority for European science and society. *The Lancet Neurology*, *15*(5), 455–532. https://doi.org/10.1016/s1474-4422(16)00062-4

Wisdom, N. M., Callahan, J. L., & Hawkins, K. A. (2011). The effects of apolipoprotein E on non-impaired cognitive functioning: A meta-analysis. *Neurobiology of Aging*, *32*(1), 63–74. https://doi.org/10.1016/j.neurobiolaging.2009.02.003

Xu, W.-L., Caracciolo, B., Wang, H.-X., Santoni, G., Winblad, B., & Fratiglioni, L. (2012). Accelerated Progression from Mild Cognitive Impairment to Dementia Among APOE 44 Carriers. *Journal of Alzheimer's Disease*, *33*(2), 507–515. https://doi.org/10.3233/JAD-2012-121369

Yesavage, J. A., & Sheikh, J. I. (1986). 9/Geriatric Depression Scale (GDS). *Clinical Gerontologist*, *5*(1-2), 165–173. https://doi.org/10.1300/j018v05n01_09

Yonelinas, A. P., Aly, M., Wang, W.-C., & Koen, J. D. (2010). Recollection and familiarity: Examining controversial assumptions and new directions. *Hippocampus*, *20*(11), 1178–1194. https://doi.org/10.1002/hipo.20864

Zeidman, P., & Maguire, E. A. (2016). Anterior hippocampus: The anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience*, *17*(3), 173–182. https://doi.org/10.1038/nrn.2015.24

Zheng, F., Cui, D., Zhang, L., Zhang, S., Zhao, Y., Liu, X., Liu, C., Li, Z., Zhang, D., Shi, L., Liu, Z., Hou, K., Lu, W., Yin, T., & Qiu, J. (2018). The Volume of Hippocampal Subfields in Relation to Decline of Memory Recall Across the Adult Lifespan.

*Frontiers in Aging Neuroscience*, *10*, 1–10. https://doi.org/10.3389/fnagi.2018.00320